PATENT ABSTRACTS OF JAPAN

(11)Publication number : *2002-073681*

(43)Date of publication of application : 12.03.2002

(51)Int.CI. G06F 17/30

(21)Application number : 2000-263240

(22)Date of filing : *28.08.2000*

(71)Applicant : HITACHI LTD

(72)Inventor : MATSUBAYASHI TADATAKA

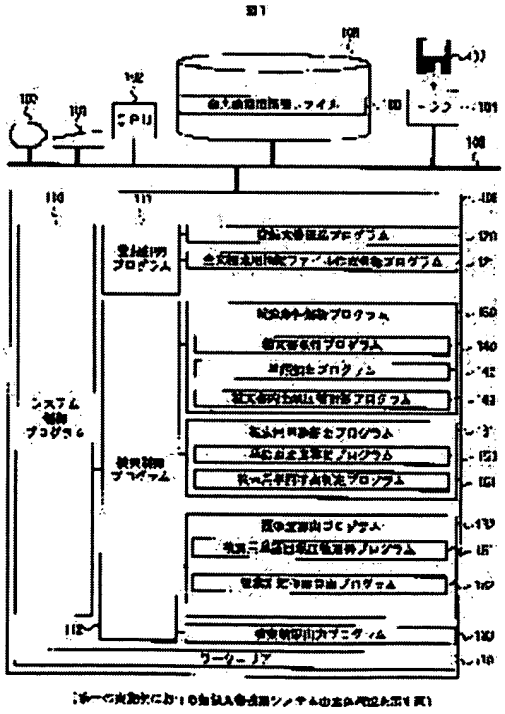
YAMAMOTO SHINYA

TADA KATSUMI

SUGAYA NATSUKO

(54) METHOD AND DEVICE FOR RETRIEVING SIMILAR DOCUMENTS AND STORAGE MEDIUM WITH PROGRAM STORED THEREIN FOR THE RETRIEVAL METHOD

(57)Abstract:



PROBLEM TO BE SOLVED: To fast retrieve the similar documents with no extreme deterioration of retrieval accuracy about a similar document retrieval method which calculates the resemblance between a master document and the registered one by referring to a full text retrieval index in a retrieval mode and without producing the feature vector of the registered document in a registration mode.

SOLUTION: This similar document retrieval method includes a full text retrieval index production process as a document registering process and also a master document feature vector production process and a resemblance calculation process as the similar document retrieval processes respectively. In such a method, a retrieval word extraction process is added after the master document feature vector production process.

LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号
特開2002－73681
（P2002－73681A）
(43)公開日　平成14年３月12日(2002.3.12)

(51)Int.Cl.⁷	識別記号	ＦＩ	ﾃｰﾏｺｰﾄﾞ(参考)
Ｇ０６Ｆ　17/30	３５０	Ｇ０６Ｆ　17/30	３５０Ｃ　５Ｂ０７５
	１７０		１７０Ａ
	３４０		３４０Ｂ

審査請求　未請求　請求項の数５　ＯＬ　（全 17 頁）

(21)出願番号　特願2000－263240(P2000－263240)

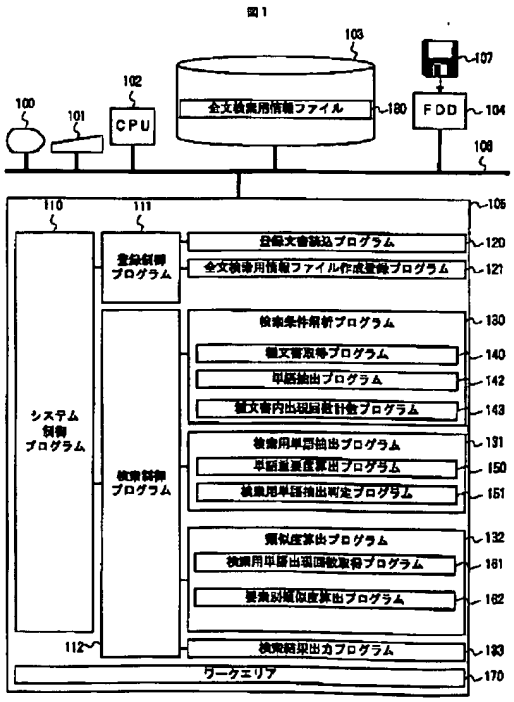
(22)出願日　平成12年８月28日(2000.8.28)

(71)出願人　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目６番地
(72)発明者　松林　忠孝
神奈川県川崎市幸区鹿島田890番地　株式会社日立製作所ビジネスソリューション開発本部内
(72)発明者　山本　伸也
神奈川県横浜市戸塚区戸塚町5030番地　株式会社日立製作所ソフトウェア事業部内
(74)代理人　100075096
弁理士　作田　康夫

最終頁に続く

(54)【発明の名称】　類似文書検索方法および装置および、類似文書検索方法のためのプログラムが記録された記憶媒体

(57)【要約】
【課題】　文書登録時に登録文書の特徴ベクトルを作成せずに、検索時に全文検索用インデクスを参照することにより、種文書と登録文書の類似度を算出する類似文書検索方法において、検索精度を極端に低下させることなく高速な類似文書検索を提供すること。
【解決手段】　文書の登録処理として全文検索用インデクス作成処理を有し、類似文書の検索処理として種文書特徴ベクトル作成処理と類似度算出処理を有する類似文書検索方法において、種文書特徴ベクトル作成処理の後に、検索用単語抽出処理を有することを特徴とする類似文書検索方法。



（第一の実施例における類似文書検索システムの全体構成を示す図）

1

【特許請求の範囲】

【請求項１】文書データベースに登録された文書あるいは文章や文字列（以下、まとめて文書と呼ぶ）から指定された文書（以下、種文書と呼ぶ）に内容が類似する文書を検索する類似文書検索方法において、文書データベースへの文書の登録処理として、登録対象とする文書の全文検索用インデクスを作成する全文検索用インデクス作成ステップと、類似文書の検索処理として、指定された種文書に含まれる文字列毎の出現回数を要素としたベクトルデータ（以下、種文書特徴ベクトルと呼ぶ）を作成する種文書特徴ベクトル作成ステップと、前記種文書特徴ベクトルの要素である文字列に対して、該種文書の中心的な内容を表す文字列をその程度（以下、文字列重要度と呼ぶ）にしたがって抽出し、該文字列重要度の降順に所定の抽出基準により類似度算出に使用する文字列（以下、検索用文字列と呼ぶ）を抽出する検索用文字列抽出ステップと、前記検索用文字列抽出ステップで抽出された検索用文字列に関して、該検索用文字列の種文書内での出現情報と、文書データベースに登録された文書（以下、登録文書と呼ぶ）内での出現情報を用いて、種文書に対する各登録文書の類似度を算出する類似度算出ステップと、前記類似度算出ステップで算出された各登録文書の種文書に対する類似度を出力する検索結果出力ステップを有することを特徴とした類似文書検索方法。

【請求項２】請求項１記載の類似文書検索方法における前記類似度算出ステップとして、前記検索用文字列抽出ステップで抽出された検索用文字列に関して、該検索用文字列の種文書内での出現回数と、登録文書内での出現回数を用いて、種文書に対する各登録文書の類似度を算出する類似度算出ステップを有することを特徴とした類似文書検索方法。

【請求項３】請求項１記載の類似文書検索方法における前記検索用文字列抽出ステップとして、前記種文書特徴ベクトル作成ステップで作成された種文書特徴ベクトルの要素である文字列について、該種文書内の出現回数を該文字列の文字列重要度とする文字列重要度算出ステップと、前記文字列重要度算出ステップで算出された文字列重要度の降順に、予め指定された個数の検索用文字列を抽出する検索用文字列判定ステップを有することを特徴とした類似文書検索方法。

【請求項４】請求項３記載の類似文書検索方法における前記検索用文字列判定ステップとして、予め指定された個数の検索用文字列を抽出する代わりに、前記文字列重要度算出ステップで算出された文字列重要度の降順に類似度算出に用いる文字列を抽出し、該文字列により種文書に対する類似度を算出し、該類似度が所定の値を超えている場合には、該文字列を検索用文字列として抽出する検索用文字列判定ステップを用いることを特徴とした類似文書検索方法。

【請求項５】請求項１記載の類似文書検索方法において、検索処理として、検索に要する時間を計測する検索処理時間測定ステップを加えるとともに、前記類似度算出ステップにおいて、上記検索処理時間測定ステップで測定された検索処理時間が所定の値を超えた場合に類似度算出処理を終了することを特徴とした類似文書検索方法。

2

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、ユーザが指定した文書に記述されている内容と類似する内容を含む文書を、文書データベースの中から検索する方法に関する。

【０００２】

【従来の技術】近年、パーソナルコンピュータやインターネット等の普及に伴い、電子化文書が爆発的に増加しており、今後も加速度的に増大していくものと予想される。このような状況において、ユーザが所望する情報を含んだ文書を高速かつ効率的に検索したいという要求が高まってきている。

【０００３】このような要求に応える技術として、ユーザが自分の所望する内容を含んだ文書（以下、種文書と呼ぶ）を例示し、その文書と類似する文書を検索する類似文書検索技術が注目されている。

【０００４】類似文書検索の方法としては、例えば「特開平１１－６６０８６」が開示されている（以下、従来技術１と呼ぶ）。

【０００５】本従来技術１では、文書データベースに対して文書を登録する際に、登録対象となる文書を全文検索するために必要な情報（従来技術１では、転置インデックスと呼んでいる。以下、全文検索用インデクスと呼ぶ。）を作成しておき、類似文書の検索時に、本全文検索用インデクスを参照することで登録済みの文書（以下、登録文書と呼ぶ）に含まれる単語の出現頻度情報を要素としてもつベクトル（以下、特徴ベクトルと呼ぶ）を作成し、これと検索条件として指定された文書（以下、種文書と呼ぶ）の特徴ベクトルとが、ベクトル空間内においてなす角度の余弦を文書間の類似度として算出する技術である。

【０００６】以下、従来技術１の処理手順を図２のＰＡＤ（Problem Analysis Diagram）図を用いて説明する。

【０００７】従来技術１では、まずステップ２００において、文書の登録処理か類似文書の検索処理かを判定し、文書の登録処理と判定された場合には全文検索用インデクス作成ステップ２１０を実行し、全文検索用インデクスを作成する。

【０００８】また、ステップ２００において類似文書の検索処理と判定された場合には、種文書特徴ベクトル生成ステップ２２０を実行し、種文書に対して特徴ベクトルを作成する。そして、全文検索用インデクスを用いた類似度算出ステップ２２１を実行し、該種文書の特徴ベクトルと登録文書の特徴ベクトルが、ベクトル空間内に

おいてなす角度の余弦を文書間の類似度として算出する。

【０００９】以上が、従来技術１の処理手順である。

【００１０】以下、図３を用いて本従来技術１の概要を説明する。

【００１１】従来技術１の文書登録処理では、まず全文検索用インデクス作成処理２１０で登録用文書１および文書２中に含まれる単語および出現位置を抽出し、全文検索用インデクス４０３を作成する。この結果、全文検索用インデクス４０３には、"構築：（文書１，５）（文書２，８）"のように記録される。ここで、"構築：（文書１，５）（文書２，８）"は、文字列"構築"が文書１の５文字目に、文書２の８文字目に出現していることを表している。

【００１２】そして、類似文書の検索処理では、検索条件で指定された種文書を抽出し、種文書特徴ベクトル生成処理２２０を通じて該種文書に対応する種文書特徴ベクトル４０６を生成する。

【００１３】次に、種文書特徴ベクトル４０６中に含まれる全ての単語に対して、前記文書登録処理で作成した全文検索用インデクス４０３を参照することで、各登録文書中の出現回数を取得する。

【００１４】ここで図４に示すように、二つのベクトルＸおよびＹの余弦は、ベクトルの対応する成分同士（例えばx(i)とy(i)）の積和値をそれぞれのベクトルの大きさで除算することにより得られることに着目する。すなわち、特定のベクトル間の内積をベクトルの組ごとに算出していくのではなく、ベクトルの要素ごとの内積成分（以下、要素別類似度と呼ぶ）を計算した後に、全ての要素における要素別類似度の総和を算出する。なお図４では、ベクトルＸのi番目の要素を"x(i)"と表し、ベクトルＸの大きさを"|Ｘ|"と表す。

【００１５】すなわち、図３において種文書特徴ベクトル４０６と登録文書の特徴ベクトルの余弦を算出するためには、種文書特徴ベクトル４０６中の全ての単語に対して、種文書と各登録文書での出現回数の積和値を各登録文書における単語毎の要素別類似度として算出し、全ての登録文書について単語毎の要素別類似度の総和をとることで算出できる。

【００１６】以下、本類似度算出方法を図５を用いて具体的に説明する。

【００１７】種文書特徴ベクトルをベクトルＸ、文書１の特徴ベクトル（以下、特徴ベクトル１と呼ぶ）をベクトルＹ、文書２の特徴ベクトル（以下、特徴ベクトル２と呼ぶ）をベクトルＺと表すとき、種文書特徴ベクトルと特徴ベクトル１および特徴ベクトル２の内積の第１成分は、それぞれ"x(1)y(1)"および"x(1)z(1)"として算出することができる。

【００１８】ここで、"x(1)"は単語１の種文書での出現回数を表しており、"y(1)"および"z(1)"はそれぞれ単語１の文書１および文書２での出現回数を表している。

【００１９】すなわち、単語１の各文書での出現回数６００は、種文書内での単語１の出現回数を計数すると共に、単語１に対応する全文検索用インデクスを参照することで取得することができる。

【００２０】以下同様に、種文書中の全ての単語に対応する全文検索用インデクスを参照することで、種文書に対する登録文書の類似度を算出することができる。

【００２１】以上が、従来技術１における類似度算出方法の具体的な説明である。

【００２２】最後に、各登録文書全体の類似度４０７を出力する。

【００２３】以上が、従来技術１の概要である。

【００２４】以上説明したように従来技術１によれば、登録文書中に含まれる単語用の全文検索用単語インデクスを予め作成しておくことで、文書検索時に登録文書の特徴ベクトルの生成を可能とし、検索条件として指定された種文書に対応する種文書特徴ベクトルとの余弦を類似度として算出することで、文書データベース中から内容の類似する文書を検索することができる。

【００２５】しかし従来技術１には、種文書から抽出された全ての単語に対して全文検索用インデクスを参照し、類似度算出に使用しているため、種文書に含まれる単語数が多いときには膨大な処理時間が必要になるということである。

【００２６】例えば、種文書中の1種類の単語に対する全文検索用インデクスを0.5秒で参照可能としても、種文書から100種類の単語が抽出されているとすると、50秒もの処理時間を要してしまうことになる。

【００２７】一方、処理時間を低減するために単純に種文書特徴ベクトルの単語を間引いてしまうと、単語の種類数を削減してしまうため種文書で重要な意味を持つ単語までもが排除される可能性があり、検索精度が極端に低下してしまう恐れがある。

【００２８】

【発明が解決しようとする課題】このような問題に対し、本発明では以下の課題を解決することを目的とする。

【００２９】すなわち本発明の課題は、文書データベースへの文書登録時に登録文書の特徴ベクトルを作成することなく、類似文書の検索時に全登録文書の特徴ベクトルを作成し、最新の単語情報を用いた類似度算出を行なう類似文書検索方法において、検索精度を確保することのできる最低限の単語数を使用することにより、高速な類似文書検索方法を実現することである。

【００３０】

【課題を解決するための手段】上記課題を解決するための、本発明に示す類似文書検索の処理手順を図７に示すＰＡＤ図に示す。

【００３１】本発明に示す類似文書検索方法は、登録処

理か研作処理かを判定する処理種別判定処理２００と、文書の登録処理として全文検索用インデクス作成処理２１０と、類似文書の検索処理として、種文書特徴ベクトル生成処理２２０と全文検索用インデクスを用いた類似度算出処理２２１を有する類似文書検索方法において、種文書特徴ベクトル生成処理２２０と全文検索用インデクスを用いた類似度算出処理２２１の間に、検索用単語抽出処理７０１を有することを特徴とする。

【００３２】すなわち、本発明による類似文書検索方法は、文書データベースへの文書登録時の全文検索用インデクス作成処理２１００として、（ステップ１）登録対象文書を読み込む登録文書読込みステップ、（ステップ２）上記登録文書読込みステップで読み込まれた登録対象文書のテキストから、全文検索用情報を抽出し、全文検索用情報ファイルに格納する全文検索用情報ファイル作成登録ステップ、と、類似文書の検索処理における種文書特徴ベクトル生成処理２２０として、（ステップ３）検索条件で指定された種文書を取得する種文書取得ステップ、（ステップ４）前記種文書読込みステップで読み込まれた種文書を解析し、種文書中に含まれる単語を抽出する種文書解析単語抽出ステップ、（ステップ５）上記種文書解析ステップで抽出された単語の出現回数を計数する種文書内出現回数計数ステップと、検索用単語抽出処理７０１として、（ステップ６）上記種文書内出現回数計数ステップで計数された各単語の出現回数に基づき、該単語の重要度を算出する単語重要度算出ステップ、（ステップ７）上記（ステップ６）で算出された各単語の重みの降順に単語を選択し、種文書自体に対する該単語の要素別類似度を算出し、該要素別類似度が所定の閾値を超える場合に、該単語を検索用単語として抽出する検索用単語判定ステップと、全文検索用インデクスを用いた類似度算出処理２２１として、（ステップ８）上記種文書特徴ベクトル生成処理２２０において、種文書から抽出された検索用単語を用いて、以下の（ステップ９）～（ステップ１０）を実行する類似度算出ステップ、（ステップ９）前記全文検索用情報ファイル作成登録ステップで作成された全文検索用情報を参照し該検索用単語の各登録文書での出現回数を取得する検索用単語出現回数取得ステップ、（ステップ１０）前記検索用単語選択ステップで選択された該検索用単語に関する前記種文書内出現回数計数ステップで取得した種文書内出現回数および前記単語出現回数取得ステップで取得した各登録文書における検索用単語出現回数を用いて種文書と登録文書の要素別類似度を算出し、各登録文書の全体の類似度に加算する要素別類似度算出ステップ、（ステップ１１）上記要素別類似度算出ステップで算出された類似度を出力する検索結果出力ステップを有する。

【００３３】上記類似文書検索方法を用いた本発明の原理について図８～図１０を用いて説明する。

【００３４】本発明の類似文書検索方法では、文書データベースへの文書登録時に（ステップ１）および（ステップ２）を実行する。

【００３５】以下、図８を用いて、文書の登録に際する処理手順の概要を説明する。

【００３６】まず、（ステップ１）で登録対象となる文書を読み込む。図８に示した例では、登録対象文書として文書１「ＬＡＮの構築と運用・保守に必要な機器を提供する。」および文書２「情報システムの構築や保守を手がけるＳＩベンダと提携する。」が登録対象文書として読み込まれる。

【００３７】次に、（ステップ２）において、上記（ステップ１）で読み込まれた登録対象文書のテキストから、全文検索用情報を抽出し、全文検索用情報ファイルに格納する。

【００３８】図８に示した例では、文書１中に含まれる"Ｌ"に対応する全文検索用情報として（文書１，１）が抽出され、全文検索用情報ファイル８０３中に格納される。なお、Ｌ（文書１，１）は、"文書１"の文字位置１に文字"Ｌ"が出現することを表す。

【００３９】また、ここで用いる全文検索用情報としては、任意の単語あるいは文字列の各登録文書での出現回数を取得することができれば、従来技術１に示したように単語インデクス方式を用いるものとしてもよいし、「特開平０８－１９４７１８」に開示されているn-gramインデクス方式を用いるものとしてもよい。

【００４０】以上が、本発明の文書登録に際する処理手順の概要である。

【００４１】次に、本発明に示した類似文書検索方法では、文書の検索時に（ステップ３）～（ステップ１１）を実行する。

【００４２】以下、図９を用いて文書の検索に際する処理手順の概要を説明する。

【００４３】まず（ステップ３）で検索条件として指定された種文書９０１「ＬＡＮシステムの構築ノウハウを武器にソリューションを展開する・・・」を読み込む。

【００４４】そして、（ステップ４）において、種文書を解析し、種文書中に含まれる単語を抽出する。ここで用いる種文書解析処理としては、従来技術１に示されるように単語辞書を参照し、単語辞書に含まれる単語を抽出される方式でもよいし、「特開平１０－１４８７２１」に開示されているように文書データベース中の統計情報を用いた単語抽出方法を用いてもよいし、種文書中に含まれるn-gramを機械的に抽出する方法であってもよいし、その他の単語抽出技術を使用しても構わない。

【００４５】図９に示した例では、この種文書解析処理の結果として、単語列９０３（ＬＡＮ，構築，ノウハウ，武器，ソリューション，展開，…）が抽出されている。

【００４６】次に、（ステップ５）において、上記（ステップ４）で抽出された単語の種文書内での出現回数を

計数し、単語と出現回数の組９０４（［ＬＡＮ，４］［構築，３］［ノウハウ，２］［武器，１］［ソリューション，２］［展開，１］…）を出力する。

【００４７】ここで、［ＬＡＮ，３］は、単語"ＬＡＮ"が3回出現しているということを表している。

【００４８】次に、（ステップ６）において、上記（ステップ５）で抽出された単語と出現回数の組９０４に対して、重要度を算出し、単語と重要度の組を出力する。この重要度の算出方法としては、例えば、種文書中の出現回数としてもよいし、データベースに登録された文書数に対する該単語の出現文書数の割合（以下、出現割合と呼ぶ）等を用いてもよい。図９に示した例では、種文書９０１中での出現回数を単語の重要度として算出し、単語重要度列９０５「［ＬＡＮ，４］［構築，３］［ソリューション，２］…を出力している。ここで、［ＬＡＮ，４］は、単語"ＬＡＮ"が重要度"４"として種文書に含まれていることを表す。

【００４９】そして、（ステップ７）において、上記（ステップ８）において算出された各単語の重要度の降順に種文書自体に対する要素別類似度を算出し、該要素別類似度が所定の閾値を超えている場合、該単語を検索用単語として抽出する。この結果として、検索用単語［ＬＡＮ，４］［構築，３］が抽出される。

【００５０】次に、（ステップ８）～（ステップ１０）において、前記（ステップ７）で取得された各単語の種文書内出現回数および前記（ステップ２）で作成された全文検索用情報ファイル８０３を参照することで、種文書に対する各登録文書の類似度を算出する。

【００５１】そして、（ステップ１１）において、類似度算出結果９０６を出力する。

【００５２】以上が、本発明の文書検索に際する処理手順の概要である。

【００５３】以下、上述した（ステップ７）により実行される検索用単語の抽出処理手順について、図１０を用いて説明する。

【００５４】まず、（ステップ７）において、前記（ステップ６）で出力された単語重要度列９０５を読み込み、重要度の降順に単語を選択する。図１０では、単語重要度列９０５「［ＬＡＮ，４］、［構築，３］、［ソリューション，２］…」から、まず［ＬＡＮ，４］を抽出している。

【００５５】そして、検索用単語"ＬＡＮ"の種文書内出現回数"４"を用いて、種文書に対する種文書の類似度の該検索用単語の要素別類似度を計算する。すなわち、登録文書として種文書と同一の文書が存在するもの（以下、仮想登録文書と呼ぶ）と仮定し、種文書特徴ベクトルと該仮想登録文書の特徴ベクトル間における該検索用単語の要素別類似度を算出し、総和を算出する。

【００５６】図１０では、検索用単語"ＬＡＮ"の種文書内出現回数"４"と仮想登録文書内出現回数"４"の積を算出し、要素別類似度"１６"を得る。

【００５７】この結果、検索用単語"ＬＡＮ"による種文書自体に対する要素別類似度は所定の閾値（本図に示した例では、５とする）を超えているため、検索用単語としてワークエリア１７０へ格納する。

【００５８】次に、［ＬＡＮ，４］の次に重要度の高い［構築，３］を選択し、種文書に対する種文書の類似度の該検索用単語の要素別類似度を計算する。この結果、要素別類似度は９となり、所定の閾値５を超えているため、検索用単語としてワークエリア１７０へ格納する。

【００５９】そして、［構築，３］の次に重要度の高い［ソリューション，２］を選択し、種文書に対する種文書の類似度の該検索用単語の要素別類似度を計算する。この結果、要素別類似度は４となり、所定の閾値を超えていないため、検索用単語として抽出せずに、終了する。

【００６０】以上が、検索用単語抽出処理手順の説明である。

【００６１】以上説明したように、文書データベースへの文書登録時に、登録文書に対する登録特徴ベクトルを作成する代わりに、全文検索用インデクスを作成しておき、類似文書の検索時には、種文書における特徴ベクトルの要素のうち種文書内での重要度の順に検索用単語を抽出し、種文書自体に対する類似度が収束するまで抽出した単語を検索用単語として使用するため、全ての単語を検索に使用する場合に比べて、検索精度を極端に落とすことなく種文書と登録文書の類似度を高速に算出することが可能となる。

【００６２】

【発明の実施の形態】以下、本発明の第一の実施例について図１を用いて説明する。

【００６３】本発明を適用した類似文書検索システムの第一例は、ディスプレイ１００、キーボード１０１、中央演算処理装置（ＣＰＵ）１０２、磁気ディスク装置１０３、フロッピディスクドライブ（ＦＤＤ）１０４、主メモリ１０５およびこれらを結ぶバス１０６から構成される。

【００６４】磁気ディスク装置１０３は二次記憶装置の一つであり、全文検索用情報ファイル１８０が格納される。

【００６５】ＦＤＤ１０４を介してフロッピディスク１０７に格納されている情報が、主メモリ１０５あるいは磁気ディスク装置１０３へ読み込まれる。

【００６６】主メモリ１０５には、システム制御プログラム１１０、登録制御プログラム１１１、検索制御プログラム１１２、登録文書読込プログラム１２０、全文検索用情報ファイル作成登録プログラム１２１、検索条件解析プログラム１３０、検索用単語抽出プログラム１３１、類似度算出プログラム１３２、検索結果出力プログラム１３３が格納されると共にワークエリア１７０が確

保される。

【００６７】検索条件解析プログラム１３０は、種文書取得プログラム１４０、単語抽出プログラム１４２および種文書内出現回数計数プログラム１４３で構成される。

【００６８】検索用単語抽出プログラム１３１は、単語重要度算出プログラム１５０および検索用単語抽出判定プログラム１５１で構成される。

【００６９】類似度算出プログラム１３２は、検索用単語出現回数取得プログラム１６１および要素別類似度算出プログラム１６２で構成される。

【００７０】登録制御プログラム１１１および検索制御プログラム１１２は、ユーザによるキーボード１０１からの指示に応じてシステム制御プログラム１１０によって起動され、それぞれ登録文書読込プログラム１２０および全文検索用情報ファイル作成登録プログラム１２１の制御と、検索条件解析プログラム１３０、検索用単語抽出プログラム１３１、類似度算出プログラム１３２および検索結果出力プログラム１３３の制御を行なう。

【００７１】なお本実施例では、キーボード１０１から入力されたコマンドにより、登録制御プログラム１１１や検索制御プログラム１１２が起動されるものとしたが、他の入力装置を介して入力されたコマンドあるいはイベントにより起動されるものであってもかまわない。

【００７２】また、これらのプログラムを磁気ディスク装置１０３、フロッピディスク１０７、MO、CD－ROM、DVD（図１には示していない）等の記憶媒体に格納し、駆動装置を介して主メモリ１０５に読み込み、CPU１０２によって実行することも可能である。

【００７３】以下、本実施例における類似文書検索システムの処理手順について説明する。

【００７４】まず、システム制御プログラム１１０の処理手順について図１１のPAD図を用いて説明する。

【００７５】システム制御プログラム１１０は、まずステップ１１００で、キーボード１０１から入力されたコマンドを解析する。

【００７６】そしてステップ１１０１で、この結果が登録実行のコマンドであると解析された場合には、ステップ１１０２で登録制御プログラム１１１を起動して、文書の登録を行なう。

【００７７】またステップ１１０１で、検索実行のコマンドであると解析された場合には、ステップ７０３で検索制御プログラム１１２を起動して、類似文書の検索を行なう。

【００７８】以上が、システム制御プログラム１１０の処理手順である。

【００７９】次に、図１１に示したステップ１１０２でシステム制御プログラム１１０により起動される登録制御プログラム１１１の処理手順について、図１２のPAD図を用いて説明する。

【００８０】登録制御プログラム１１１では、まずステップ１２００において登録文書読込プログラム１２０を起動し、登録対象として指定された文書（以下、登録対象文書と呼ぶ）を読み込み、ワークエリア１７０に格納する。

【００８１】次に、ステップ１２０１において、全文検索用情報ファイル作成登録プログラム１２１を起動し、ワークエリア１７０に格納されている登録文書に対応する全文検索用情報を作成し、全文検索用情報ファイル１８０へ格納する。

【００８２】以上が、登録制御プログラム１１１の処理手順である。

【００８３】次に、図１１に示したステップ１１０３でシステム制御プログラム１１０により起動される検索制御プログラム１１２の処理手順について、図１３のPAD図を用いて説明する。

【００８４】検索制御プログラム１１２は、まずステップ１３００において、検索条件解析プログラム１３０を起動し、種文書から単語を抽出する。

【００８５】次にステップ１３０１において、検索用単語抽出プログラム１３１を起動し、上記ステップ１３００において種文書から抽出された単語の重要度を算出し、所定の条件に基づいて重要度の高い単語を検索用単語として抽出する。

【００８６】そしてステップ１３０２において、類似度算出プログラム１３２を起動し、上記ステップ１３０１において種文書から抽出された検索用単語の出現情報を用いて、種文書に対する各登録文書の類似度を算出する。

【００８７】そしてステップ１３０３において、検索結果出力プログラム１３３を起動し、上記ステップ１３０２で算出された類似度算出結果を検索結果として出力する。

【００８８】ここで、検索結果の出力先は、ディスプレイ１００に表示するものとしてもよいし、ワークエリア１７０や磁気ディスク１０３上に格納するものとしてもよい。また、類似度算出結果をディスプレイ１００に出力する場合には、類似度の降順に出力するものとしてもよいし、文書に付与された管理番号の昇順あるいは降順に出力するものとしてもよい。

【００８９】以上が検索制御プログラム１１２の処理手順である。

【００９０】次に、図１３に示したステップ１３００で検索制御プログラム１１２により起動される検索条件解析プログラム１３０の処理手順について、図１４のPAD図を用いて説明する。

【００９１】検索条件解析プログラム１３０は、まずステップ１４００において、種文書取得プログラム１４０を起動し、検索条件で指定された種文書を抽出し、ワークエリア１７０に格納する。

【００９２】次にステップ１４０２において、単語抽出プログラム１４２を起動し、ワークエリア１７０に格納された種文書から単語を抽出する。

【００９３】そしてステップ１４０３において、種文書内出現回数計数プログラム１４３を起動し、ステップ１４０２で抽出された単語について、種文書内での出現回数を計数し、ワークエリア１７０に格納する。

【００９４】以上が検索条件解析プログラム１３０の処理手順である。

【００９５】次に、図１３に示したステップ１３０１で検索制御プログラム１１２により起動される検索用単語抽出プログラム１３１の処理手順について、図１５のＰＡＤ図を用いて説明する。

【００９６】検索用単語抽出プログラム１３１は、まずステップ１５００において、単語重要度算出プログラム１５１を起動し、所定の算出式に基づきワークエリア１７０に格納された単語の重要度を算出し、ワークエリア１７０に格納する。

【００９７】次に、前記ステップ１５００でワークエリア１７０に格納された全ての単語に対して、ステップ１５０２～１５０５を繰り返し実行する（ステップ１５０１）。

【００９８】まず、ステップ１５０２において、ワークエリア１７０に格納されている単語を重要度の降順に取得する。

【００９９】次に、ステップ１５０３において、検索用単語抽出判定プログラム１５１を起動し、種文書の要素別類似度を算出する。

【０１００】そして、ステップ１５０４において、種文書の要素別類似度が、所定の閾値を超えているかを判定し、超えている場合にはステップ１５０５を、越えていない場合には繰り返し処理を終了する。

【０１０１】そして、ステップ１５０５において、該単語を検索用単語としてワークエリア１７０に格納する。

【０１０２】以上が検索用単語抽出プログラム１３１の処理手順である。

【０１０３】なお、上述のステップ１５０２における各単語の要素別類似度の算出方法は、従来技術１に示されるように、各単語の種文書での出現回数を用いて算出してもよいし、後述するように、該単語の文書データベースでの出現文書数等の統計情報を用いるものでもよいし、さらには、文書内での出現位置情報を考慮することもできる。

【０１０４】次に、図１３に示したステップ１３０２で検索制御プログラム１１２により起動される類似度算出プログラム１３２の処理手順について、図１６のＰＡＤ図を用いて説明する。

【０１０５】類似度算出プログラム１３２は、ワークエリア１７０に格納された全ての検索用単語に対して、ステップ１６０２～１６０３を繰り返し実行する（ステップ１６０１）。

【０１０６】ステップ１６０２では、検索用単語出現回数取得プログラム１６１を起動し、検索用単語に対応する全文検索用情報ファイル１８０を参照して、各登録文書内での出現回数を取得し、ワークエリア１７０に格納する。

【０１０７】次にステップ１６０３において、要素別類似度算出プログラム１６２を起動し、ワークエリア１７０に格納された検索用単語の種文書内出現回数および登録文書内出現回数を用いて、所定の算出式により種文書に対する登録文書の要素別類似度を算出し、登録文書全体の類似度に加算する。

【０１０８】以上が類似度算出プログラム１３２の処理手順である。

【０１０９】以上が、本発明の第一の実施形態である。

【０１１０】なお、本実施例では、検索条件解析プログラム１３０により種文書から単語が抽出されるものとしたが、単語の代わりにn-gramが抽出されるものとしてもよい。この場合、検索用単語抽出プログラム１３１により処理される単位もn-gramとなる。

【０１１１】また、検索用単語抽出プログラム１３１のステップ１５０４では、ステップ１５０３で算出された種文書の要素別類似度が所定の閾値を超えるか否かを判定するものとしたが、要素別類似度ではなく類似度の総和が所定の閾値を超えているかを判定するものとしてもよいし、さらには、種文書から抽出された全ての単語における要素別類似度の総和に対する類似度の算出割合が所定の閾値を超えているかを判定するものとしてもよい。

【０１１２】また、本実施例では種文書に対する各登録文書の類似度の算出には、単語の出現回数を直接用いたが、さらにこれを種文書や登録文書の文書の長さ等により正規化してもよいことは明らかであろう。

【０１１３】以上説明したように、本発明の第一の実施形態によれば、種文書に対する要素別類似度の値を目安にして類似度算出に使用する検索用単語数を削減しているため、種文書に対する類似度算出結果が収束する必要最低限の検索で処理を終了させることができる。

【０１１４】この結果として、検索精度を極端に低下させることなく検索用単語数を削減することができ、高速な類似文書検索を実現することができるようになる。

【０１１５】なお、本実施例では、登録対象文書や種文書を文書としたが、文章あるいは文字列であっても構わないことは明らかであろう。

【０１１６】また、以上説明した本発明の第一の実施例における検索用単語抽出プログラム１３１では、種文書の要素別類似度の値を目安にして検索用単語を削減するものとしたが、予め指定された数の検索用単語を抽出するものとしてもよい。この場合の検索用単語数の設定方法としては、予め用意したテストパターンを用いて所定

の時間以内に検索が終了するように検索用単語数を決定するものとしてもよい。

【０１１７】次に本発明の第二の実施例について図１７を用いて説明する。

【０１１８】本発明を適用した類似文書検索システムの第二の実施例は、種文書から抽出された単語の重要度を算出する際に、文書データベースに蓄積された登録文書の統計情報を利用するものである。

【０１１９】本方法によれば、第一の実施例における単語重要度算出プログラム１５０による単語重要度算出の際に、種文書内の出現情報だけでなく文書データベース全体での出現情報を利用することができ、文書データベース内で頻繁に出現する単語の重要度を調整することが可能となり、第一の実施例に比べ高精度に単語重要度を算出できるようになる。

【０１２０】本実施例は、第一の実施例（図１）とほぼ同様の構成を取るが、単語重要度算出プログラム１５０の構成が異なり、図１７に示すように統計情報参照プログラム１７００が加わる。

【０１２１】以下、第一の実施例と異なる単語重要度算出プログラム１５０ａの処理手順について図１８を用いて説明する。

【０１２２】単語重要度算出プログラム１５０ａは、まずステップ１８００において、統計情報参照プログラム１７００を起動し、全文検索用情報ファイル１８０を参照することにより、種文書から抽出された各単語の文書データベースにおける出現文書数を該単語の統計情報として取得する。

【０１２３】なお、全文検索用情報ファイル１８０から該単語の出現文書数の取得は、図８に示した全文検索用情報ファイル８０３として示したように全文検索用情報ファイル１８０には各単語の文書番号および出現位置が格納されていることを利用し、該単語の異なる文書番号を計数することで実現することができる。

【０１２４】そして、ステップ１８０１において、種文書から抽出された各単語の重要度を、該単語の種文書内出現回数および文書データベースにおける統計情報を用いて算出し、ワークエリア１７０に格納する。

【０１２５】以上が、単語重要度算出プログラム１５０ａの処理手順である。

【０１２６】なお、本実施例における単語重要度算出式としては、例えばＴＦ・ＩＤＦ（Text Frequency, Inverted Documents Frequency）法を用いるものとしてもよい。

【０１２７】以上が本発明の第二の実施例である。

【０１２８】以上説明したように、本発明の第二の実施例における類似文書検索システムを用いることにより、文書データベース内で頻繁に出現する単語（以下、頻出単語と呼ぶ）を考慮した単語重要度を算出できるようになる。すなわち、頻出単語の単語重要度を低く、希少な単語の単語重要度を高く設定することで、種文書の特徴を表す単語を優先的に選択することが可能となり、高精度な類似文書検索を実現することができるようになる。

【０１２９】次に、本発明の第三の実施例について図１９を用いて説明する。

【０１３０】本発明を適用した類似文書検索システムの第三の実施例は、第二の実施例と同様に種文書から抽出された単語の重要度を算出する際に、文書データベースに蓄積された登録文書の統計情報を利用するものであるが、統計情報の取得に統計情報ファイル１９００を利用する点が異なる。

【０１３１】本方法によれば、第二の実施例における単語重要度算出の際に参照する統計情報取得を高速に行なうことができるようになる。

【０１３２】本実施例は、第二の実施例（図１７）とほぼ同様の構成を取るが、登録制御プログラム１１１の構成が異なり、図１９に示すように統計情報ファイル作成登録プログラム１９００が加わる。また、磁気ディスク装置１０３には統計情報ファイル１９１０が格納される。前記単語重要度算出プログラム１５０ａのステップ１８００では、種文書から抽出された各単語の文書データベースにおける統計情報を取得する際に、全文検索用情報ファイル１８０を参照する代わりに、図１９に示す統計情報ファイル１９１０を参照するようになる。

【０１３３】以下、第二の実施例と異なる登録制御プログラム１１１ａの処理手順について図２０を用いて説明する。

【０１３４】登録制御プログラム１１１ａでは、まずステップ１２００において登録文書読込プログラム１２０を起動し、登録対象として指定された文書を読み込み、ワークエリア１７０に格納する。

【０１３５】次に、ステップ１２０１において、全文検索用情報ファイル作成登録プログラム１２１を起動し、ワークエリア１７０に格納されている登録文書に対応する全文検索用情報を作成し、全文検索用情報ファイル１８０へ格納する。

【０１３６】次に、ステップ２０００において、統計情報ファイル作成登録プログラム１９００を起動し、ワークエリア１７０に格納されている登録文書に対応する統計情報を作成し、統計情報ファイル１９１０へ格納する。

【０１３７】以上が、登録制御プログラム１１１の処理手順である。

【０１３８】図２１に統計情報ファイル作成登録プログラム１９００により作成される統計情報ファイル１９１０の例を示す。

【０１３９】本図に示した統計情報ファイル１９１０には、管理番号２１００、単語２１０１および出現文書数２１０２が格納される。

【０１４０】本図に示した例では、管理番号"０"の領域

に、単語"ＬＡ"が格納され、該単語の出現文書数が"１"であるというように格納されることを示している。

【０１４１】なお、図２１に示した例では、統計情報ファイル１９００を表形式で格納されるものとしたが、単語と出現文書数が取得できる形式であればどのような形式であってもかまわない。例えば、トライ形式で格納されるものとしてもかまわないし、全文検索用情報ファイル１８０の先頭領域に格納しておくものとしてもかまわない。

【０１４２】以上が、本発明の第三の実施例である。

【０１４３】以上説明したように本発明の第三の実施例によれば、種文書から抽出された各単語の統計情報を取得に、文書登録処理時に予め作成された統計情報ファイルを参照することにより、全文検索用情報を参照して異なる出現文書番号の個数を計数する必要がなくなり、高速に統計情報を取得することができるようになる。これにより、第二の実施例に比べ高速な類似文書検索を実現できるようになる。

【０１４４】次に本発明の第四の実施例について図２２を用いて説明する。

【０１４５】本発明を適用した類似文書検索システムの第四の実施例は、種文書から抽出された各単語の統計情報を近似して利用するものである。

【０１４６】本方法によれば、統計情報の精度を極端に低下させることなく、第三の実施例における統計情報ファイル１９１０に格納される統計情報の容量を削減することができるようになる。

【０１４７】本実施例は、第三の実施例（図１９）とほぼ同様の構成を取るが、統計情報参照プログラム１７００の構成が異なり、近似統計情報算出プログラム２２００が加わる。

【０１４８】以下、第三の実施例と異なる統計情報参照プログラム１７００ｂの処理手順について図２３を用いて説明する。

【０１４９】統計情報参照プログラム１７００ｂは、種文書から抽出された全ての単語についてステップ２３０１～２３０４を繰り返し実行する（ステップ２３００）。

【０１５０】ステップ２３０１では、統計情報ファイル１９１０を参照し、該単語に対応する統計情報が格納されているかを確認する。

【０１５１】そして、該単語が統計情報ファイル１９１０中に格納されている場合にはステップ２３０３を実行し、格納されていない場合にはステップ２３０４を実行する（ステップ２３０２）。

【０１５２】ステップ２３０３では、統計情報ファイル１９１０を参照し、該単語の統計情報を取得する。

【０１５３】また、ステップ２３０４では、近似統計情報算出プログラム２２００を起動し、該単語の近似統計情報を算出する。

【０１５４】以上が、統計情報参照プログラム１７００ｂの処理手順である。

【０１５５】次に、近似統計情報算出プログラム２２００の処理手順について図２４を用いて具体的に説明する。

【０１５６】本図に示した例では、まずステップ２３０１において、統計情報を取得する対象となる単語２４００"ＬＡＮ"対して、統計情報ファイル１９１０を参照する。

【０１５７】ここでは、統計情報ファイル１９１０には"ＬＡＮ"が格納されていないため、ステップ２３０４を実行する。

【０１５８】ステップ２３０４では、単語２４００"ＬＡＮ"の構成要素である"ＬＡ"と"ＡＮ"の統計情報をそれぞれ取得し、これらの出現文書数のうち少ない値を"ＬＡＮ"の統計情報として設定する。

【０１５９】本図に示した例では、"ＬＡ"の統計情報２４０１に格納された出現文書数"８０７"と、"ＡＮ"の統計情報２４０２に格納された出現文書数"１５１２"とを比較し、この結果として"ＬＡＮ"の統計情報２４０３として値の小さい"ＬＡ"の出現文書数"８０７"を格納する（２４１０）。

【０１６０】これは、単語"ＬＡＮ"の構成要素"ＬＡ"と"ＡＮ"の出現文書数が異なる場合、"ＬＡＮ"の出現文書数は各構成要素よりも多くなることはありえないという性質を利用するものである。すなわち、単語"ＬＡＮ"の出現文書数としては、本来"ＬＡＮ"そのものの出現文書数を用いるべきであるが、単語"ＬＡＮ"の構成要素である"ＬＡ"あるいは"ＡＮ"のうち、出現文書数の少ない値を近似した出現文書数として参照するものである。

【０１６１】以上が近似統計情報算出プログラム２２００の具体的な処理手順である。

【０１６２】以上が本発明の第四の実施例である。

【０１６３】以上説明したように、本発明の第四の実施例における類似文書検索システムを用いることにより、全ての単語の出現文書数を統計情報ファイルへ格納する必要がなくなるため、第三の実施例に比べ、統計情報ファイルの容量を削減することができるようになる。

【０１６４】以上説明したように、本発明の第一の実施例から第四の実施例における類似文書検索システムでは、種文書の類似度を算出し、これに基づいて検索用単語数を調整しているため、検索精度を確保しながら高速に類似文書検索を実現することができる。

【０１６５】次に、本発明の第五の実施例について図２５を用いて説明する。

【０１６６】本発明を適用した類似文書検索システムの第五の実施例は、所定の検索時間で検索結果を出力するものである。

【０１６７】本方法によれば、ユーザは所定の検索時間で検索結果を取得できるため、検索条件で指定した種文

書が検索目的に合致しているかをストレスなく判断できるようになる。

【０１６８】本実施例は、第一の実施例（図１）とほぼ同様の構成を取るが、類似度算出プログラム１３２の構成が異なり、検索処理時間計測プログラム２５００が加わる。

【０１６９】以下、第一の実施例と異なる類似度算出プログラム１３２ｂの処理手順を図２６のＰＡＤ図を用いて説明する。

【０１７０】類似度算出プログラム１３２ｂは、ステップ２６００において、検索処理時間計測プログラム２５００を起動し、検索処理時間の計測を開始する。

【０１７１】次に、ワークエリア１７０に格納された全ての検索用単語に対して、検索処理時間が所定の値（以下、検索制限時間と呼ぶ）以下ならば、ステップ１６０２、１６０３および２６０２を繰り返し実行する（ステップ２６０１）。

【０１７２】ステップ１６０２では、検索用単語出現回数取得プログラム１６１を起動し、検索用単語に対応する全文検索用情報ファイル１８０を参照して、各登録文書内での出現回数を取得し、ワークエリア１７０に格納する。

【０１７３】次にステップ１６０３において、要素別類似度算出プログラム１６２を起動し、ワークエリア１７０に格納された検索用単語の種文書内出現回数および登録文書内出現回数を用いて、所定の算出式により種文書に対する登録文書の要素別類似度を算出し、登録文書全体の類似度に加算する。

【０１７４】そして、ステップ２６０２において、検索処理時間計測プログラム２５００を起動し、検索処理時間の経過時間を測定し、検索処理時間を算出する。

【０１７５】以上が類似度算出プログラム１３２ｂの処理手順である。

【０１７６】以上が本発明の第五の実施形態である。

【０１７７】なお、本実施例のステップ２６０１における検索制限時間は、検索実行時に検索条件として指定するものとしてもよいし、システム設定値として予め設定しておくものとしてもよい。

【０１７８】また、本実施例では、検索制限時間を設定するものとしたが、設定値によっては少数の検索用単語しか用いられない場合も考えられるため、検索精度を保つための最小限の検索用単語数を設定できるようにしてもよい。この場合は、検索処理時間が検索制限時間を上回ったとしても、指定された最小限の検索用単語数までは類似検索を繰り返すことになる。

【０１７９】さらに、本実施例では、検索処理時間計測プログラム２５００を用いて類似度算出処理に要する時間を計測するものとしたが、検索処理自体を計測するものとしてもよい。この場合、図２６に示したステップ２６００で検索時間の計測を開始するのではなく、検索制御プログラム１１２により検索条件解析プログラム１３０が起動される前に、検索処理時間計測プログラム２５００を起動し、検索処理時間の測定を開始すればよい。

【０１８０】以上説明したように本発明の第五の実施例における類似文書検索システムでは、検索に要する時間に基づいて検索用単語数を調整するため、所定の処理時間で検索結果を取得することができるようになる。

【０１８１】この結果として、ユーザは検索終了時間を予測することができるようになる。

【０１８２】なお、第一の実施例から第四の実施例で説明した種文書の類似度を目安に検索を終了する類似文書検索システムと第五の実施例で説明した検索時間を目安に検索を終了する類似文書検索システムを検索実行時あるいはシステム定義で切り替えて使用することも可能である。

【０１８３】次に、本発明の第六の実施例について図２７を用いて説明する。

【０１８４】本発明を適用した類似文書検索システムの第六の実施例は、種文書から抽出された単語から検索に使用される検索用単語から、検索時間を推定し、長大な時間を要する場合にはユーザに確認を求めるものである。

【０１８５】本方法によれば、第一の実施例から第四の実施例で説明した類似文書検索システムにおける検索用単語抽出条件では検索に長大な時間を要する場合、事前に検索を取りやめることができるため、ユーザは不用意に待たされることがなくなる。

【０１８６】本実施例は、第一の実施例（図１）とほぼ同様の構成を取るが、検索用単語抽出プログラム１３１の構成が異なり、図２７に示すように検索時間推定確認プログラム２７００が加わる。

【０１８７】以下、第一の実施例と異なる検索用単語抽出プログラム１３１ｂの処理手順を図２８のＰＡＤ図を用いて説明する。

【０１８８】検索用単語抽出プログラム１３１では、まずステップ１５００において、単語重要度算出プログラム１５１を起動し、所定の算出式に基づきワークエリア１７０に格納された単語の重要度を算出し、ワークエリア１７０に格納する。

【０１８９】次に、前記ステップ１５００でワークエリア１７０に格納された全ての単語に対して、ステップ１５０２～１５０５を繰り返し実行する（ステップ１５０１）。

【０１９０】まず、ステップ１５０２において、ワークエリア１７０に格納されている単語を重要度の降順に取得する。

【０１９１】次に、ステップ１５０３において、検索用単語抽出判定プログラム１５１を起動し、種文書の要素別類似度を算出する。

【０１９２】そして、ステップ１５０４において、種文

書の要素別類似度が、所定の閾値を超えているかを判定し、超えている場合にはステップ１５０５を、越えていない場合には繰り返し処理を終了する。

【０１９３】そして、ステップ１５０５において、該単語を検索用単語としてワークエリア１７０に格納する。

【０１９４】次に、ステップ２８００において、ワークエリア１７０に格納された検索用単語から検索時間を推定し、推定された検索時間（以下、推定検索時間と呼ぶ）が所定の値（指定検索時間）を超える場合には、検索の継続を確認するメッセージを表示し、ユーザの確認を受ける。この確認メッセージとしては、例えば図６に示したように、継続ボタン２９０１およびキャンセルボタン２９０１を有するメッセージ２９００を表示するものであってもよい。

【０１９５】以上が検索用単語抽出プログラム１３１ｂの処理手順である。

【０１９６】なお、上記ステップ２８００における指定検索時間としては、検索条件として指定するものとしてもよいし、システム定義として予め指定されるものとしてもよいし、あるいはいくつかのテストパターンの結果から自動的に設定されるものとしてもよい。

【０１９７】また、上記ステップ２８００における検索時間の推定方法としては、該検索用単語の出現文書数から推定するものとしてもよいし、該検索用単語に対応する全文検索用情報ファイル１８０のサイズから推定するものとしてもよい。あるいは、いくつかのテストパターンを用いてひとつの検索用単語に要する平均時間を計測しておき、該平均時間を用いて検索時間を推定するものとしてもよい。

【０１９８】以上説明したように、本実施例に示した類似文書検索システムでは、抽出された検索用単語から検索時間を推定し、推定検索時間が予め指定された時間を超える場合には検索用単語の抽出条件を調整することが可能となるため、ユーザは不用意に待たされることがなくなる。

【０１９９】

【発明の効果】以上説明したように、本発明では、種文書の類似度を目安に検索用単語数を設定しているため、類似度算出に使用する検索用単語数を削減することができる。これにより、検索精度を確保することのできる高速な類似文書検索を実現することができる。

【図面の簡単な説明】

【図１】本発明の第一の実施例における類似文書検索システムの全体構成を示す図である。

【図２】従来技術１の処理手順を説明するＰＡＤ図である。

【図３】従来技術１の概要を説明する図である。

【図４】従来技術１の類似度算出方式の考え方を説明する図である。

【図５】従来技術１の類似度算出方式の考え方を説明する図である。

【図６】本発明の第六の実施例における検索時間推定確認プログラム２７００による確認メッセージの例である。

【図７】本発明の処理手順を説明するＰＡＤ図である。

【図８】本発明の登録処理の概要を説明する図である。

【図９】本発明の検索処理の概要を説明する図である。

【図１０】本発明の検索用単語抽出処理の概要を説明する図である。

【図１１】本発明の第一の実施例におけるシステム制御プログラム１１０の処理手順を説明する図である。

【図１２】本発明の第一の実施例における登録制御プログラム１１１の処理手順を説明する図である。

【図１３】本発明の第一の実施例における検索制御プログラム１１２の処理手順を説明するＰＡＤ図である。

【図１４】本発明の第一の実施例における検索条件解析プログラム１３０の処理手順を説明するＰＡＤ図である。

【図１５】本発明の第一の実施例における検索用単語抽出プログラム１３１の処理手順を説明するＰＡＤ図である。

【図１６】本発明の第一の実施例における類似度算出プログラム１３２の処理手順を説明するＰＡＤ図である。

【図１７】本発明の第二の実施例における単語重要度算出プログラム１５０ａの構成を示す図である。

【図１８】本発明の第三の実施例における単語重要度算出プログラム１５０ａの処理手順を説明するＰＡＤ図である。

【図１９】本発明の第三の実施例における登録制御プログラム１１１ａの構成図である。

【図２０】本発明の第三の実施例における登録制御プログラム１１１ａの処理手順を示すＰＡＤ図である。

【図２１】本発明の第三の実施例における統計情報ファイル１９１０の例である。

【図２２】本発明の第四の実施例における統計情報参照プログラム１７００ｂの構成を示す図である。

【図２３】本発明の第四の実施例における統計情報参照プログラム１７００ｂの処理手順を説明するＰＡＤ図である。

【図２４】本発明の第四の実施例における近似統計情報の算出方法を説明する図である。

【図２５】本発明の第五の実施例における類似度算出プログラム１３２ｂの構成を示す図である。

【図２６】本発明の第五の実施例における類似度算出プログラム１３２ｂの処理手順を説明するＰＡＤ図である。

【図２７】本発明の第六の実施例における検索用単語抽出プログラム１３１ｂの構成を示す図である。

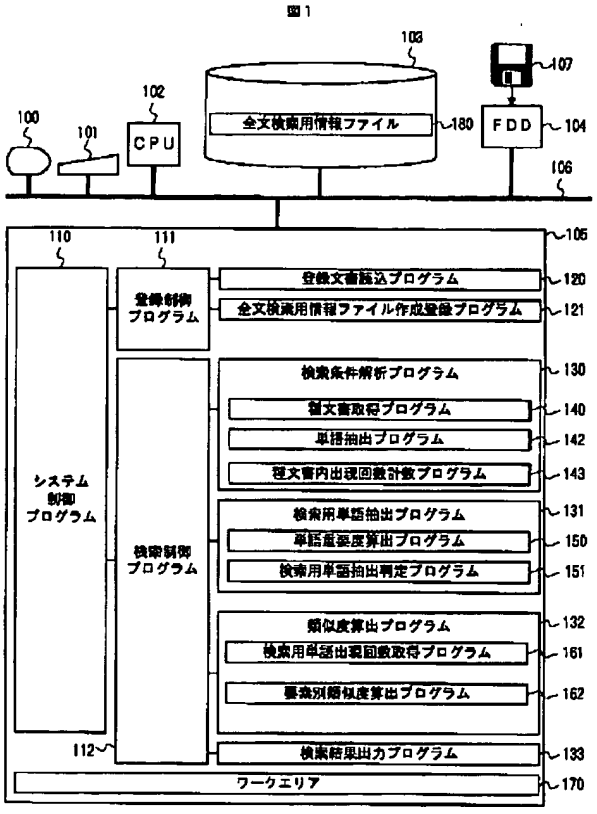
【図２８】本発明の第六の実施例における検索用単語抽出プログラム１３１ｂの処理手順を説明するＰＡＤ図で

ある。

【符号の説明】

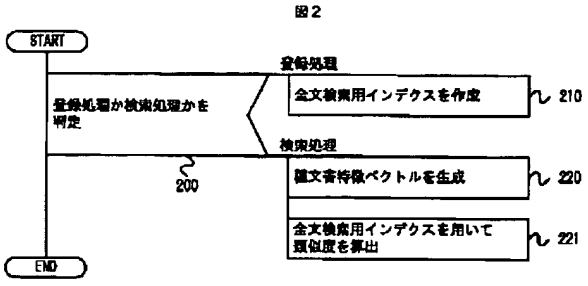
１００…ディスプレイ、１０１…キーボード、１０２…中央演算処理装置（ＣＰＵ）、１０３…磁気ディスク装置、１０４…フロッピディスクドライブ（ＦＤＤ）、１０５…主メモリ、１０６…バス、１０７…フロッピディスク、１１０…システム制御プログラム、１１１…登録制御プログラム、１１２…検索制御プログラム、１２０…登録文書読込プログラム、１２１…全文検索用情報ファイル作成登録プログラム、１３０…検索条件解析プロ＊グラム、１３１…検索用単語抽出プログラム、１３２…類似度算出プログラム、１３３…検索結果出力プログラム、１４０…種文書取得プログラム、１４２…単語抽出プログラム、１４３…種文書内出現回数計数プログラム、１５０…単語重要度算出プログラム、１５１…検索用単語抽出判定プログラム、１６１…検索用単語出現回数取得プログラム、１６２…要素別類似度算出プログラム、１７０…ワークエリア、１８０…全文検索用情報ファイル。

【図１】



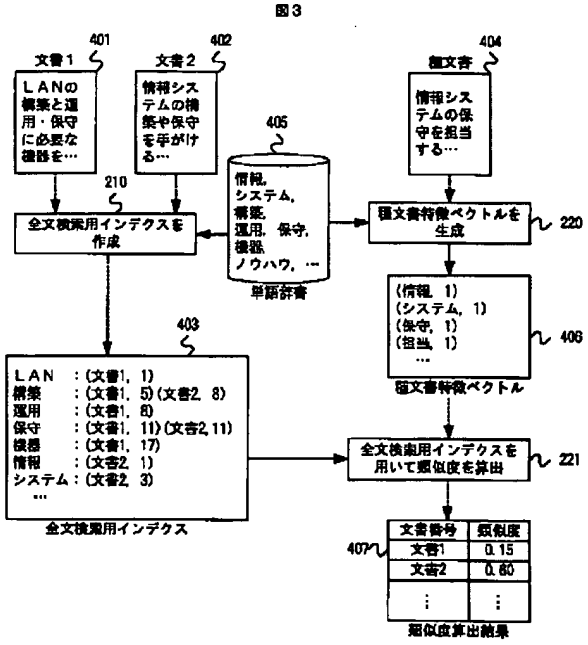
（第一の実施例における類似文書検索システムの全体構成を示す図）

【図２】



（従来技術１の処理手順）

【図３】



（従来技術１の概要を説明する図）

【図４】

図４

$$X(x(1),\ x(2),\ x(3),\ \cdots,\ x(i),\ \cdots)$$

$$Y(y(1),\ y(2),\ y(3),\ \cdots,\ y(i),\ \cdots)$$

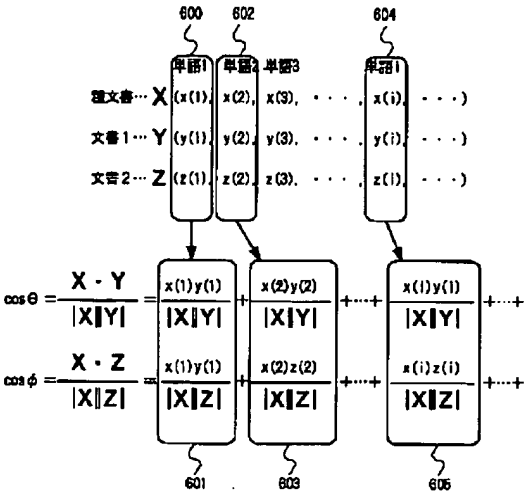
$$\cos\theta = \frac{X \cdot Y}{|X||Y|} = \frac{x(1)y(1)+x(2)y(2)+\cdots+x(i)y(i)+\cdots}{|X||Y|}$$

$$= \frac{x(1)y(1)}{|X||Y|} + \frac{x(2)y(2)}{|X||Y|} + \cdots + \frac{x(i)y(i)}{|X||Y|} + \cdots +$$

（従来技術１の類似度算出方式の考え方を説明する図）

【図５】

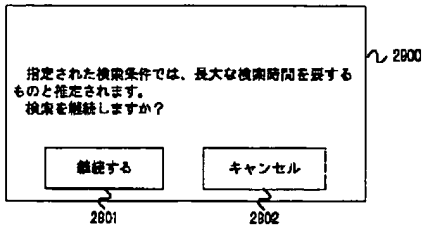
図５



（従来技術１の類似度算出方式の考え方を説明する図）

【図６】

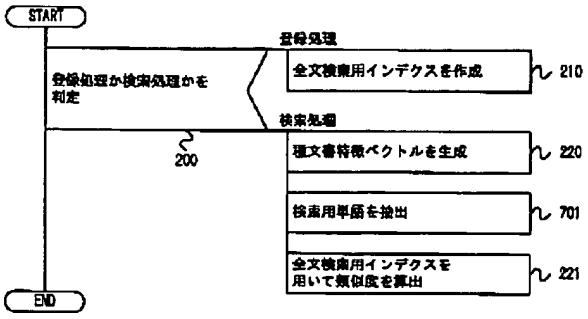
図６



（検索時間推定確認プログラム２７００による確認メッセージの例）

【図７】

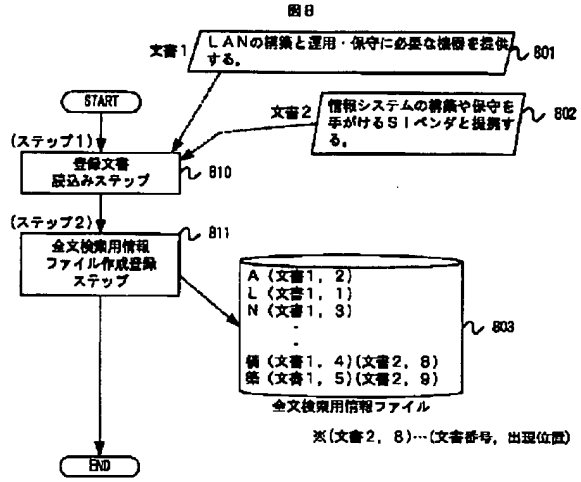
図７



（本発明の処理手順）

【図８】

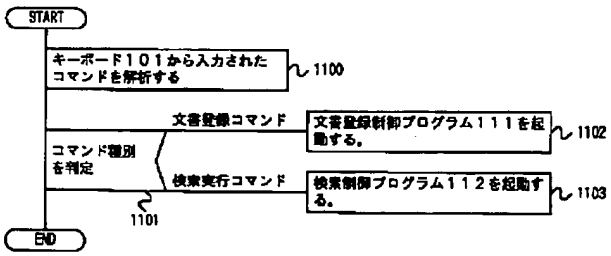
図８



（本発明の登録処理の概要を説明する図）

【図１１】

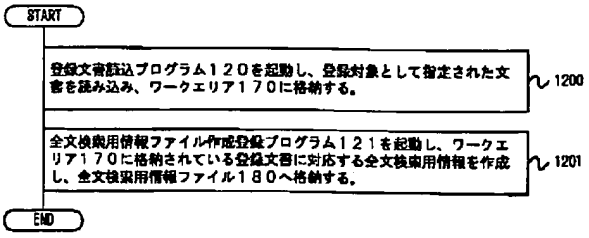
図１１



（システム制御プログラム１１０の処理手順）

【図１２】

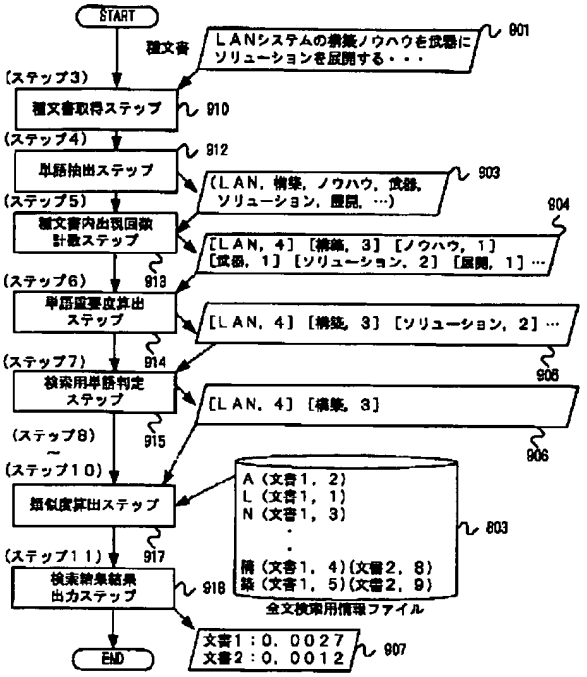
図１２



（登録制御プログラム１１１の処理手順）

【図９】

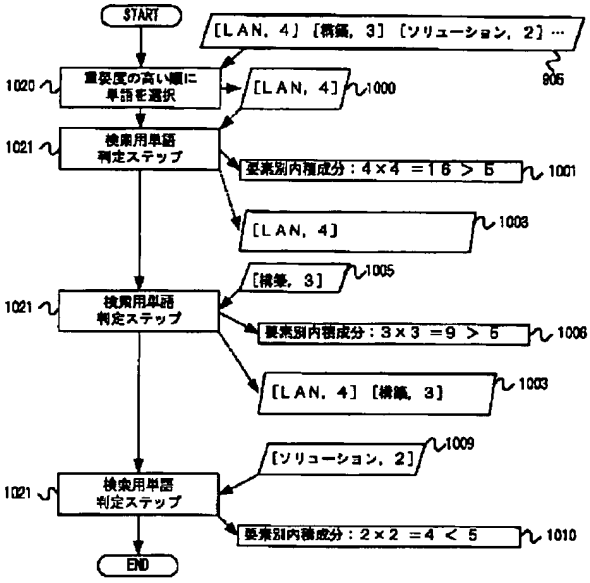
図９



（本発明の検索処理の概要を説明する図）

【図１０】

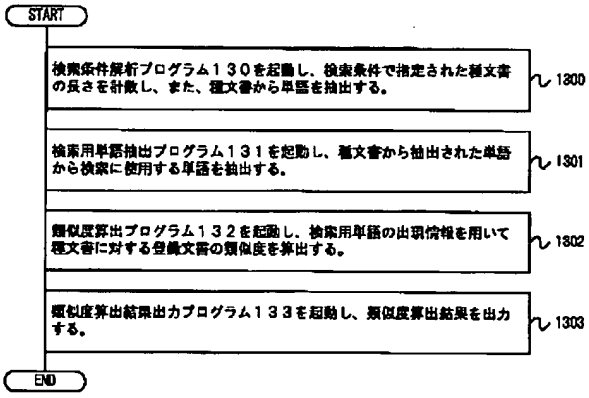
図１０



（本発明の検索用単語抽出処理の概要を説明する図）

【図１３】

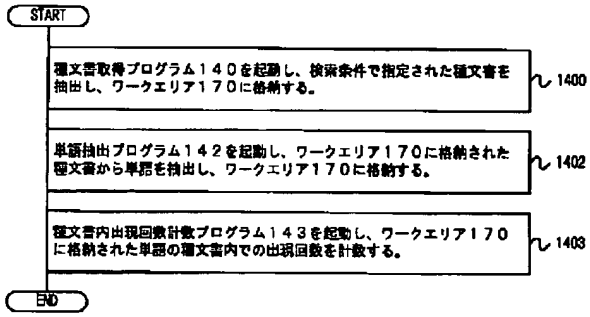
図１３



（検索制御プログラム１１２の処理手順）

【図１４】

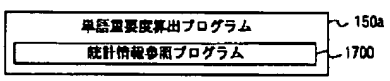
図１４



（検索条件解析プログラム１３０の処理手順）

【図１７】

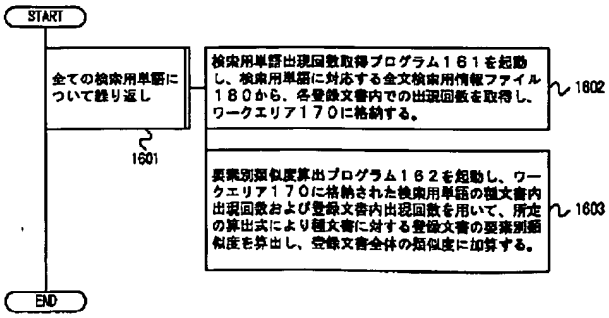
図１７



（本発明の第二の実施例における単語重要度算出プログラム１５０ａの構成を示す図）

【図１６】

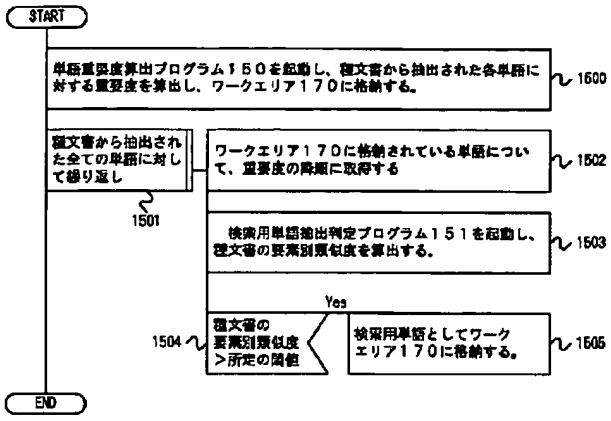
図１６



（類似度算出プログラム１３２の処理手順）

【図１５】

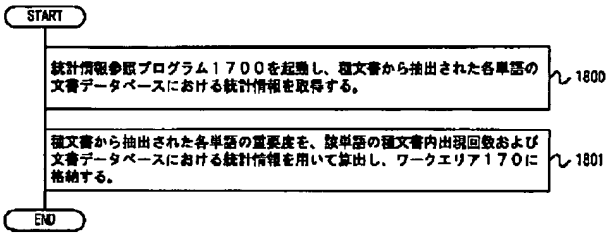
図１５



（検索用単語抽出プログラム１３１の処理手順）

【図１８】

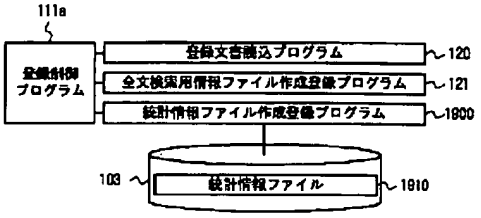
図１８



（単語重要度算出プログラム１５０ａの処理手順）

【図１９】

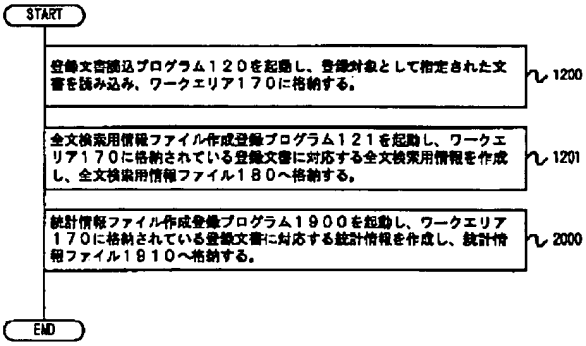
図１９



（本発明の第三の実施例における登録制御プログラム１１１ａの構成図）

【図２０】

図２０



（登録制御プログラム１１１ａの処理手順）

【図２１】

図２１

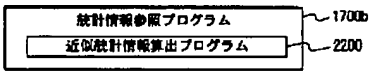
2100 管理番号	2101 単語	2102 出現文書数
0	ＳＡ	1
1	ＡＮ	1
2	構築	2
3	運用	1
4	保守	2
5	機器	1
6	情報	1
7	シス	1
8	ステ	1
9	テム	1
⋮	⋮	⋮

1910

（統計情報ファイル１９１０の例）

【図２２】

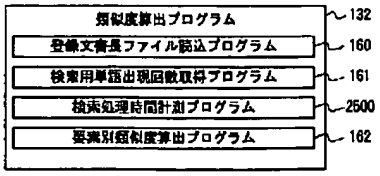
図２２



（本発明の第四の実施例における統計情報参照プログラム１７００ｂの構成を示す図）

【図２５】

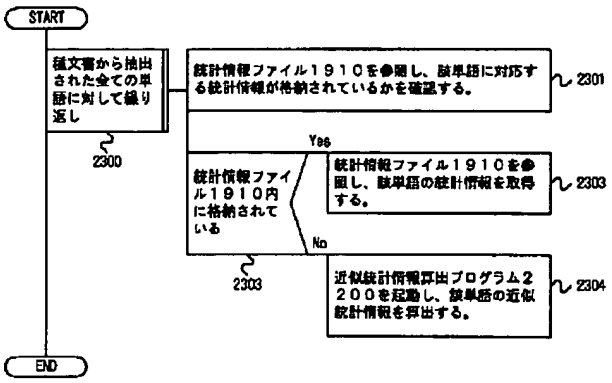
図２５



（本発明の第五の実施例における類似度算出プログラム１３２ｂの構成を示す図）

【図２３】

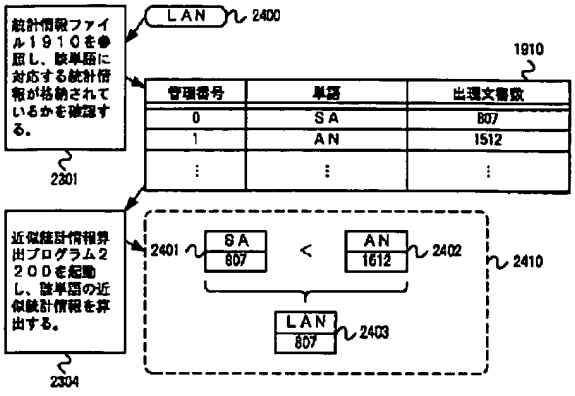
図２３



（統計情報参照プログラム１７００ｂの処理手順を説明する図）

【図２４】

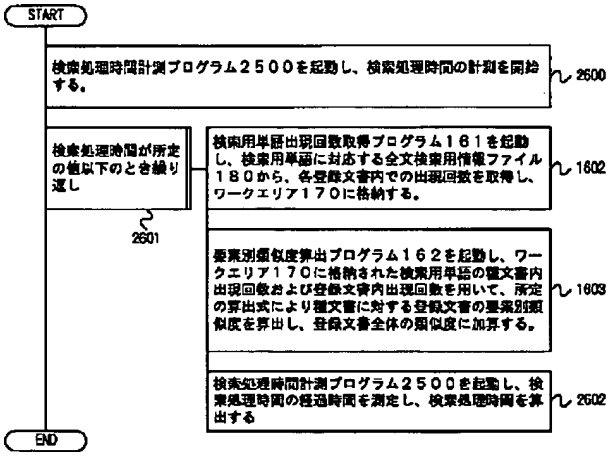
図２４



（近似統計情報の算出方法を説明する図）

【図２６】

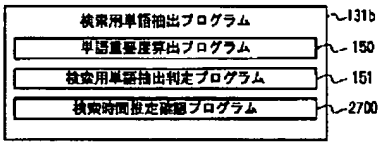
図２６



（類似度算出プログラム１３２ｂの処理手順）

【図２７】

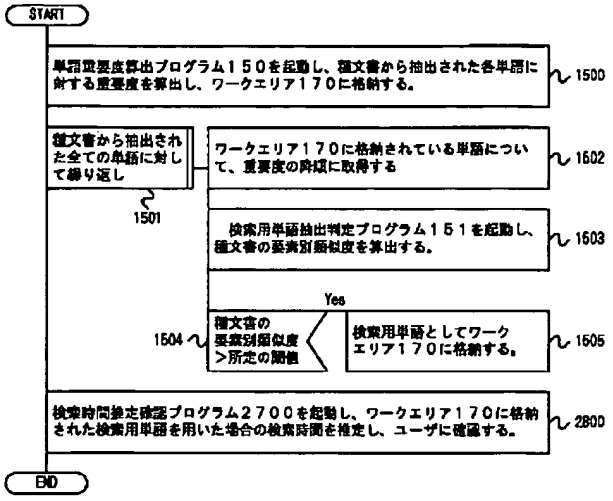
図２７



（本発明の第六の実施例における検索用単語抽出プログラム１３１ｂの構成を示す図）

【図２８】

図２８



（検索用単語抽出プログラム１３１ｂの処理手順）

フロントページの続き

(72)発明者　多田　勝己
神奈川県川崎市幸区鹿島田890番地　株式会社日立製作所ビジネスソリューション開発本部内

(72)発明者　菅谷　奈津子
神奈川県川崎市幸区鹿島田890番地　株式会社日立製作所ビジネスソリューション開発本部内

Ｆターム(参考)　5B075 ND03 NK32 PP02 PQ02 PQ74 PR04 PR06 PR08 QM08

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

DETAILED DESCRIPTION

[Detailed Description of the Invention]
[0001]
[Field of the Invention] This invention relates to the approach of searching a document including the contents described by the document specified by a user, and similar contents out of a document database.
[0002]
[Description of the Prior Art] In recent years, with the spread of a personal computer, the Internet, etc., the electronic document is increasing explosively and is expected to continue to increase at an increasing tempo. In such a situation, the high speed and the demand of wanting to search efficiently have been increasing the document including the information for which a user asks.
[0003] As a technique which meets such a demand, the document (it is hereafter called a seed document) with which the user included the contents for which he asks is illustrated, and the similar document-retrieval technique of searching the document and a similar document attracts attention.
[0004] As the approach of a similar document retrieval, "JP,11-66086,A" is indicated, for example (it is hereafter called the conventional technique 1).
[0005] Information required [in case a document is registered to a document database] of this conventional technique 1 in order to carry out the full-text search of the document used as the candidate for registration (with the conventional technique 1, it is called the transposition index.) Hereafter, it is called the index for full-text searches. It creates. At the time of retrieval of a similar document The vector which has as an element the frequency-of-occurrence information on the word contained in a registered document (it is hereafter called a registration document) by referring to the index for these full-text searches The feature vector of the document (it is hereafter called a seed document) which created (it is hereafter called a feature vector) and was specified as this and retrieval conditions is the technique which computes the cosine of the include angle made in vector space as similarity between documents.
[0006] Hereafter, the procedure of the conventional technique 1 is explained using the PAD (Problem Analysis Diagram) Fig. of drawing 2 .
[0007] With the conventional technique 1, first, in step 200, when registration processing of a document or retrieval processing of a similar document is judged and it is judged with registration processing of a document, the index creation step 210 for full-text searches is performed, and the index for full-text searches is created.
[0008] Moreover, when judged with retrieval processing of a similar document in step 200, the seed document feature-vector generation step 220 is performed, and a feature vector is created to a seed document. And the similarity calculation step 221 using the index for full-text searches is performed, and the feature vector of this seed document and the feature vector of a registration document compute the cosine of the include angle made in vector space as similarity between documents.
[0009] The above is the procedure of the conventional technique 1.
[0010] Hereafter, the outline of this conventional technique 1 is explained using drawing 3 .

[0011] In document registration processing of the conventional technique 1, the word and appearance location which are first included in the document 1 for registration and a document 2 by the index creation processing 210 for full-text searches are extracted, and the index 403 for full-text searches is created. Consequently, it is recorded on the index 403 for full-text searches like "construction: (documents 1 and 5) (documents 2 and 8)." Here, "construction:(documents 1 and 5) (documents 2 and 8) "character string" construction" means having appeared in the 8th character of a document 2 to the 5th character of a document 1.
[0012] And in retrieval processing of a similar document, the seed document specified on retrieval conditions is extracted, and the seed document feature vector 406 corresponding to this seed document is generated through the seed document feature-vector generation processing 220.
[0013] Next, the count of an appearance in each registration document is acquired by referring to the index 403 for full-text searches created by said document registration processing to all the words contained in the seed document feature vector 406.
[0014] The cosine of two vectors X and Y notes being obtained by doing the division of the sum-of-products value of the components (for example, x (i) and y (i)) to which a vector corresponds by each magnitude of a vector here at drawing 4 so that it may be shown. That is, after calculating the inner product component (it is hereafter called the similarity according to element) for every element of a vector rather than computing the inner product between specific vectors for every group of a vector, total of the similarity according to element in all elements is computed. In addition, the i-th element of Vector X is expressed in drawing 4 as "x (i)", and the magnitude of Vector X is expressed in it as "|X|."
[0015] That is, in order to compute the cosine of the seed document feature vector 406 and the feature vector of a registration document in drawing 3 , to all the words in the seed document feature vector 406, the sum-of-products value of the count of an appearance in a seed document and each registration document is computed as similarity according to element for every [in each registration document] word, and it can compute by taking total of the similarity according to element for every word about all registration documents.
[0016] Hereafter, this similarity calculation approach is concretely explained using drawing 5 .
[0017] When a seed document feature vector is expressed as Vector Z, the 1st component of the inner product of a seed document feature vector, a feature vector 1, and a feature vector 2 can compute [a feature vector] Vector Y and the feature vector (it is hereafter called a feature vector 2) of a document 2 for Vector X and the feature vector (it is hereafter called a feature vector 1) of a document 1 as "x(1) y (1)" and "x(1) z (1)", respectively.
[0018] Here, "x (1)" expresses the count of an appearance in the seed document of a word 1, and "y (1)" and "z (1)" express the count of an appearance in the document 1 and document 2 of a word 1, respectively.
[0019] That is, the count 600 of an appearance in each document of a word 1 is acquirable by referring to the index for full-text searches corresponding to a word 1 while carrying out counting of the count of an appearance of the word 1 within a seed document.
[0020] The similarity of a registration document to a seed document is computable like the following by referring to the index for full-text searches corresponding to all the words in a seed document.
[0021] The above is concrete explanation of the similarity calculation approach in the conventional technique 1.
[0022] Finally, the similarity 407 of each whole registration document is outputted.
[0023] The above is the outline of the conventional technique 1.
[0024] As explained above, generation of the feature vector of a registration document is enabled at the time of a document retrieval by creating beforehand the word index for full-text searches for words contained in a registration document according to the conventional technique 1, and the document with which the contents are similar out of a document database can be searched with computing as similarity a cosine with the seed document feature vector corresponding to the seed document specified as retrieval conditions.
[0025] However, since it is used for similarity calculation with reference to the index for full-text

searches to all the words extracted from the seed document, when there are many words contained in a seed document, I hear that the huge processing time is needed for the conventional technique 1, and it is in it.
[0026] For example, supposing 100 kinds of words are extracted also as reference being possible in 0.5 seconds after a seed document in the index for full-text searches to one kind of word in a seed document, the processing time as long as 50 seconds will be required.
[0027] On the other hand, in order to reduce the processing time, when the word of a seed document feature vector is thinned out simply, since the number of classes of a word is reduced, even the word which has important semantics by the seed document may be eliminated, and there is a possibility that retrieval precision may fall extremely.
[0028]
[Problem(s) to be Solved by the Invention] By this invention, it aims at solving the following technical problems to such a problem.
[0029] That is, the technical problem of this invention is realizing the high-speed similar document-retrieval approach by creating the feature vector of all registration documents at the time of retrieval of a similar document, and using the minimum number of words which can secure retrieval precision in the similar document-retrieval approach of performing similarity calculation using the newest word information, without creating the feature vector of a registration document at the time of the document registration to a document database.
[0030]
[Means for Solving the Problem] The procedure of a similar document retrieval shown in this invention for solving the above-mentioned technical problem is shown in the PAD diagram shown in drawing 7 .
[0031] The processing classification judging processing 200 in which the similar document-retrieval approach shown in this invention judges registration processing or the Kensaku processing, In the similar document-retrieval approach of having the seed document feature-vector generation processing 220 and the similarity calculation processing 221 in which the index for full-text searches was used, as registration processing of a document as the index creation processing 210 for full-text searches, and retrieval processing of a similar document It is characterized by having the word extract processing 701 for retrieval between the seed document feature-vector generation processing 220 and the similarity calculation processing 221 using the index for full-text searches.
[0032] Namely, the similar document-retrieval approach by this invention As index creation processing 2100 for full-text searches at the time of the document registration to a document database (Step 1) From the text of the document for registration read at the registration document read in step and the above-mentioned (step 2) registration document read in step which read the document for registration The information file creation registration step for full-text searches which extracts the information for full-text searches and is stored in the information file for full-text searches, As seed document feature-vector generation processing 220 in retrieval processing of a similar document The seed document acquisition step which acquires the seed document specified on retrieval conditions, (Step 3) The seed document read at said seed document read in step is analyzed. (Step 4) the count of the appearance in a seed document which carries out counting of the count of an appearance of the word extracted at the seed document analysis word extract step and the above-mentioned (step 5) seed document analysis step which extract the word contained in a seed document -- counting -- with a step It is based on the count of an appearance of each word by which counting was carried out at the step. as the word extract processing 701 for retrieval -- the above-mentioned (step 6) count of the appearance in a seed document -- counting -- A word is chosen as the descending order of the weight of each word computed by the word significance calculation step which computes the significance of this word, and the above (step 7) (step 6). The word judging step for retrieval which extracts this word as a word for retrieval when the similarity according to element of this word to the seed document itself is computed and this similarity according to element exceeds a predetermined threshold, As similarity calculation processing 221 using the index for full-text searches, it sets to the above-mentioned (step 8) seed document feature-vector generation processing 220. The similarity calculation step which performs the following - (step 9) (step

10) using the word for retrieval extracted from the seed document, The count acquisition step for retrieval of a word appearance which acquires the count of an appearance in each registration document of this word for retrieval with reference to the information for full-text searches created at said information file creation registration step for full-text searches, (Step 9) (Step 10) said count of the appearance in a seed document about this word for retrieval chosen at said word selection step for retrieval -- counting -- the count for retrieval of a word appearance in each registration document acquired at the count of the appearance in a seed document acquired at the step, and said count acquisition step of a word appearance It has the retrieval result output step which outputs the similarity which computed the similarity according to element of a seed document and a registration document by having used, and was computed at the similarity calculation step classified by element added to the similarity of each whole registration document, and the above-mentioned (step 11) similarity calculation step classified by element.

[0033] The principle of this invention using the above-mentioned similar document-retrieval approach is explained using drawing 8 - drawing 10 .

[0034] the similar document-retrieval approach of this invention -- the time of the document registration to a document database -- and (step 1) (step 2) it performs.

[0035] Hereafter, the outline of procedure of facing registration of a document is explained using drawing 8 .

[0036] First, the document which serves as a candidate for registration at (step 1) is read. In the example shown in drawing 8 , a document 1 "a device required for construction of LAN, and employment and maintenance is offered." and a document 2 "it ties up with SI vendor which deals with construction and maintenance of an information system." are read as a document for registration as a document for registration.

[0037] Next, in (step 2), from the text of the document for registration read above (step 1), the information for full-text searches is extracted and it stores in the information file for full-text searches.

[0038] In the example shown in drawing 8 , (documents 1 and 1) are extracted as information for full-text searches corresponding to "L" contained in a document 1, and it is stored in the information file 803 for full-text searches. In addition, L (documents 1 and 1) means that alphabetic character"L" appears in the character position 1 of "a document 1."

[0039] Moreover, as information for full-text searches that it uses here, if the word of arbitration or the count of an appearance in each registration document of a character string is acquirable, as shown in the conventional technique 1, it is good also as a thing using a word index method, and good also as a thing using the n-gram index method currently indicated by "JP,08-194718,A."

[0040] The above is the outline of procedure of facing document registration of this invention.

[0041] Next, by the similar document-retrieval approach shown in this invention, - (step 3) (step 11) is performed at the time of retrieval of a document.

[0042] Hereafter, the outline of procedure of facing retrieval of a document using drawing 9 is explained.

[0043] The seed document 901 "a solution is developed for the construction know-how of a LAN system to arms ..." specified as retrieval conditions first (step 3) is read.

[0044] And in (step 4), a seed document is analyzed and the word contained in a seed document is extracted. As seed document analysis processing in which it uses here, as shown in the conventional technique 1, a word dictionary is referred to. The method from which the word contained in a word dictionary is extracted may be used, may use the word extract approach using the statistical information in a document database as indicated by "JP,10-148721,A", and You may be the approach of extracting mechanically n-gram contained in a seed document, and it does not matter even if it uses other word extract techniques.

[0045] In the example shown in drawing 9 , the word train 903 (LAN, construction, know-how, arms, a solution, expansion, --) is extracted as a result of this seed document analysis processing.

[0046] Next, in (step 5), counting of the count of an appearance within the seed document of the word extracted above (step 4) is carried out, and the group 904 ([LAN, 4], 3 [[construction and 3]], 2

[[know-how and 2]], 1 [[arms and 1]], 2 [[a solution and 2]], [expansion and 1] --) of a word and the count of an appearance is outputted.
[0047] Here, [LAN, 3] mean that word"LAN" has appeared 3 times.
[0048] Next, in (step 6), to the group 904 of the word extracted above (step 5) and the count of an appearance, significance is computed and the group of a word and significance is outputted. the number of appearance documents of this word [as opposed to the number of documents good as a count of an appearance in a seed document for example, which carried out and was registered into the database as the calculation approach of this significance] -- comparatively (the following and an appearance -- it calls comparatively) -- etc. -- you may use. In the example shown in drawing 9 , the count of an appearance in the inside of the seed document 901 is computed as a significance of a word, and word significance train 905" [LAN, 4], [construction and 3], and [solution and 2] -- are outputted. Here, [LAN, 4] mean that word"LAN" is contained in a seed document as significance "4."
[0049] And in (step 7), when the similarity according to element to the seed document itself is computed in descending order of the significance of each word computed in the above (step 8) and this similarity according to element is over the predetermined threshold, this word is extracted as a word for retrieval. As this result, the word for retrieval [LAN, 4], and [construction and 3] are extracted.
[0050] Next, in - (step 8) (step 10), the similarity of each registration document to a seed document is computed by referring to the information file 803 for full-text searches created with the count of the appearance in a seed document of each word acquired above (step 7), and the above (step 2).
[0051] And the similarity calculation result 906 is outputted in (step 11).
[0052] The above is the outline of procedure of facing the document retrieval of this invention.
[0053] the following -- having mentioned above (step 7) -- the extract procedure of the word for retrieval performed is explained using drawing 10 .
[0054] First, in (step 7), the word significance train 905 outputted above (step 6) is read, and a word is chosen as the descending order of significance. In drawing 10 , [LAN, 4] are first extracted from the word significance train "[LAN, 4], [construction and 3], [a solution and 2] --" 905.
[0055] And the similarity according to element of this word for retrieval of the similarity of a seed document to a seed document is calculated using word "count of the appearance in kind document of LAN"" 4for retrieval." That is, it assumes that it is that (it is hereafter called a virtual registration document) in which the document same as a registration document as a seed document exists, the similarity according to element of this word for retrieval between a seed document feature vector and the feature vector of this virtual registration document is computed, and total is computed.
[0056] drawing 10 -- the object for retrieval -- "LAN "count of the appearance in seed document" 4" and count of the appearance in virtual registration document" of word 4 -- " -- a product -- computing -- the similarity according to element -- "16" is obtained.
[0057] Consequently, since the similarity according to element to the seed document by word"LAN" for retrieval itself is over the predetermined threshold (referred to as 5 in the example shown in this Fig.), it is stored in a work area 170 as a word for retrieval.
[0058] Next, [construction and 3] with a high significance are chosen as the degree of [LAN, 4], and the similarity according to element of this word for retrieval of the similarity of a seed document to a seed document is calculated. Consequently, since the similarity according to element was set to 9 and is over the predetermined threshold 5, it is stored in a work area 170 as a word for retrieval.
[0059] And [a solution and 2] with a high significance are chosen as the degree of [construction and 3], and the similarity according to element of this word for retrieval of the similarity of a seed document to a seed document is calculated. Consequently, since the similarity according to element is set to 4 and is not over the predetermined threshold, it is ended, without extracting as a word for retrieval.
[0060] The above is explanation of the word extract procedure for retrieval.
[0061] Instead of creating the registration feature vector to a registration document at the time of the document registration to a document database, as explained above The index for full-text searches is created. At the time of retrieval of a similar document In order to use the word which extracted the word for retrieval in order of the significance within a seed document among the elements of the feature

vector in a seed document, and was extracted until it was completed by the similarity to the seed document itself as a word for retrieval, Compared with the case where all words are used for retrieval, it becomes possible to compute the similarity of a seed document and a registration document at a high speed, without dropping retrieval precision extremely.
[0062]
[Embodiment of the Invention] Hereafter, the first example of this invention is explained using drawing 1 .
[0063] The first example of the similar document-retrieval system which applied this invention consists of buses 106 which connect a display 100, a keyboard 101, arithmetic and program control (CPU) 102, a magnetic disk drive 103, the floppy disk drive (FDD) 104, main memory 105, and these.
[0064] A magnetic disk drive 103 is one of the secondary storages, and the information file 180 for full-text searches is stored.
[0065] The information stored in the floppy disk 107 through FDD104 is read into main memory 105 or a magnetic disk drive 103.
[0066] While a system control program 110, the registration control program 111, the retrieval control program 112, the registration document read in program 120, the information file creation registration program 121 for full-text searches, the retrieval condition analyzer 130, the word extract program 131 for retrieval, the similarity calculation program 132, and the retrieval result output program 133 are stored, a work area 170 is secured to main memory 105.
[0067] the retrieval condition analyzer 130 -- the seed document acquisition program 140, the word extract program 142, and the count of the appearance in a seed document -- counting -- it consists of programs 143.
[0068] The word extract program 131 for retrieval consists of a word significance calculation program 150 and a word extract judging program 151 for retrieval.
[0069] The similarity calculation program 132 consists of a count acquisition program 161 for retrieval of a word appearance, and a similarity calculation program 162 classified by element.
[0070] The registration control program 111 and the retrieval control program 112 are started by the system control program 110 according to the directions from the keyboard 101 by the user, and perform control of the registration document read in program 120 and the information file creation registration program 121 for full-text searches, and control of the retrieval condition analyzer 130, the word extract program 131 for retrieval, the similarity calculation program 132, and the retrieval result output program 133, respectively.
[0071] In addition, although the registration control program 111 and the retrieval control program 112 shall be started in this example by the command inputted from the keyboard 101, the command or event inputted through other input devices may start.
[0072] Moreover, it is also possible to store these programs in storages, such as a magnetic disk drive 103, a floppy disk 107, MO, CD-ROM, and DVD (not shown in drawing 1), to read into main memory 105 through a driving gear, and to perform by CPU102.
[0073] Hereafter, the procedure of the similar document-retrieval system in this example is explained.
[0074] First, the procedure of a system control program 110 is explained using the PAD diagram of drawing 11 .
[0075] A system control program 110 is step 1100 first, and analyzes the command inputted from the keyboard 101.
[0076] And at step 1101, when this result is analyzed as it is the command of registration activation, the registration control program 111 is started at step 1102, and a document is registered.
[0077] Moreover, when it is analyzed that it is the command of retrieval activation, the retrieval control program 112 is started at step 703, and a similar document is searched with step 1101.
[0078] The above is the procedure of a system control program 110.
[0079] Next, the procedure of the registration control program 111 started by the system control program 110 at step 1102 shown in drawing 11 is explained using the PAD diagram of drawing 12 .
[0080] In the registration control program 111, in step 1200, the registration document read in program

120 is started first, the document (it is hereafter called the document for registration) specified as a candidate for registration is read, and it stores in a work area 170.
[0081] Next, in step 1201, the information file creation registration program 121 for full-text searches is started, the information for full-text searches corresponding to the registration document stored in the work area 170 is created, and it stores in the information file 180 for full-text searches.
[0082] The above is the procedure of the registration control program 111.
[0083] Next, the procedure of the retrieval control program 112 started by the system control program 110 at step 1103 shown in drawing 11 is explained using the PAD diagram of drawing 13 .
[0084] First, in step 1300, the retrieval control program 112 starts the retrieval condition analyzer 130, and extracts a word from a seed document.
[0085] Next, in step 1301, the word extract program 131 for retrieval is started, the significance of the word extracted from the seed document in the above-mentioned step 1300 is computed, and a word with a high significance is extracted as a word for retrieval based on predetermined conditions.
[0086] And in step 1302, the similarity calculation program 132 is started and the similarity of each registration document to a seed document is computed using the appearance information on the word for retrieval extracted from the seed document in the above-mentioned step 1301.
[0087] And in step 1303, the retrieval result output program 133 is started and the similarity calculation result computed at the above-mentioned step 1302 is outputted as a retrieval result.
[0088] Here, the output destination change of a retrieval result is good also as what is displayed on a display 100, and good also as what is stored on a work area 170 or a magnetic disk 103. Moreover, when outputting a similarity calculation result to a display 100, it is good also as what is outputted to the descending order of similarity, and good also as what is outputted to the ascending order or descending order of a management number given to the document.
[0089] The above is the procedure of the retrieval control program 112.
[0090] Next, the procedure of the retrieval condition analyzer 130 started by the retrieval control program 112 at step 1300 shown in drawing 13 is explained using the PAD diagram of drawing 14 .
[0091] First, the retrieval condition analyzer 130 starts the seed document acquisition program 140, extracts the seed document specified on retrieval conditions in step 1400, and stores it in a work area 170.
[0092] Next, in step 1402, the word extract program 142 is started and a word is extracted from the seed document stored in the work area 170.
[0093] and the step 1403 -- setting -- the count of the appearance in a seed document -- counting -- a program 143 is started, and about the word extracted at step 1402, counting of the count of an appearance within a seed document is carried out, and it stores in a work area 170.
[0094] The above is the procedure of the retrieval condition analyzer 130.
[0095] Next, the procedure of the word extract program 131 for retrieval started by the retrieval control program 112 at step 1301 shown in drawing 13 is explained using the PAD diagram of drawing 15 .
[0096] First, the word extract program 131 for retrieval starts the word significance calculation program 151, computes the significance of the word stored in the work area 170 based on the predetermined formula in step 1500, and stores it in a work area 170.
[0097] Next, steps 1502-1505 are repeated and performed to all the words stored in the work area 170 at said step 1500 (step 1501).
[0098] First, in step 1502, the word stored in the work area 170 is acquired in descending order of significance.
[0099] Next, in step 1503, the word extract judging program 151 for retrieval is started, and the similarity according to element of a seed document is computed.
[0100] And in step 1504, when the similarity according to element of a seed document judges and is over whether it is over the predetermined threshold and it is not over step 1505, repeat processing is ended.
[0101] And in step 1505, it stores in a work area 170 by making this word into the word for retrieval.
[0102] The above is the procedure of the word extract program 131 for retrieval.

[0103] In addition, as shown in the conventional technique 1, the calculation approach of the similarity according to element of each word in the above-mentioned step 1502 may compute using the count of an appearance in the seed document of each word, and it can also take into consideration the appearance positional information within a document further using statistical information, such as the number of appearance documents in the document database of this word, so that it may mention later.
[0104] Next, the procedure of the similarity calculation program 132 started by the retrieval control program 112 at step 1302 shown in drawing 13 is explained using the PAD diagram of drawing 16 .
[0105] The similarity calculation program 132 repeats and performs steps 1602-1603 to all the words for retrieval stored in the work area 170 (step 1601).
[0106] At step 1602, the count acquisition program 161 for retrieval of a word appearance is started, the count of an appearance within each registration document is acquired with reference to the information file 180 for full-text searches corresponding to the word for retrieval, and it stores in a work area 170.
[0107] Next, in step 1603, the similarity calculation program 162 classified by element is started, the similarity according to element of the registration document to a seed document is computed by the predetermined formula using the count of the appearance in a seed document of the word for retrieval stored in the work area 170, and the count of the appearance in a registration document, and it adds to the similarity of the whole registration document.
[0108] The above is the procedure of the similarity calculation program 132.
[0109] The above is the first operation gestalt of this invention.
[0110] In addition, in this example, although a word shall be extracted from a seed document by the retrieval condition analyzer 130, it is good also as that from which n-gram is extracted instead of a word. In this case, the unit processed by the word extract program 131 for retrieval also serves as n-gram.
[0111] Moreover, although the similarity according to element of the seed document computed at step 1503 shall judge whether a predetermined threshold is exceeded at step 1504 of the word extract program 131 for retrieval It is good also as what judges whether total of the similarity instead of the similarity according to element is over the predetermined threshold, and good also as what judges whether the calculation rate of the similarity to total of the similarity according to element in all the words extracted from the seed document is over the predetermined threshold further.
[0112] Moreover, although the count of an appearance of a word was directly used for calculation of the similarity of each registration document to a seed document in this example, probably, it will be clear that this may be further normalized with the die length of the document of a seed document or a registration document etc.
[0113] Since the number for retrieval of words which follows the value of the similarity according to element to a seed document as a guide, and is used for similarity calculation is reduced according to the first operation gestalt of this invention as explained above, processing can be terminated by the necessary minimum retrieval which the similarity calculation result of a seed document converges.
[0114] The number for retrieval of words can be reduced as this result, without reducing retrieval precision extremely, and a high-speed similar document retrieval can be realized now.
[0115] In addition, in this example, although the document for registration and the seed document were used as the document, probably, it will be clear that you may be a text or a character string.
[0116] Moreover, although the value of the similarity according to element of a seed document shall be followed as a guide and the words for retrieval shall be reduced in the word extract program 131 for retrieval in the first example of this invention explained above, it is good also as what extracts a number of words for retrieval specified beforehand. It is good also as what determines the number for retrieval of words that retrieval is completed within predetermined time amount using the test pattern prepared beforehand as the setting approach of the number for retrieval of words in this case.
[0117] Next, the second example of this invention is explained using drawing 17 .
[0118] In case the second example of the similar document-retrieval system which applied this invention computes the significance of the word extracted from the seed document, it uses the statistical information of the registration document accumulated in the document database.
[0119] According to this approach, in the case of the word significance calculation by the word

significance calculation program 150 in the first example, the appearance information not only on the appearance information in a seed document but the whole document database can be used, it becomes possible to adjust the significance of the word which appears frequently within a document database, and word significance can be computed now with high precision compared with the first example.
[0120] Although this example takes the almost same configuration as the first example (drawing 1), the configurations of the word significance calculation program 150 differ, and as shown in drawing 17 , the statistical information reference program 1700 is added.
[0121] Hereafter, the procedure of different word significance calculation program 150a from the first example is explained using drawing 18 .
[0122] Word significance calculation program 150a acquires first the number of appearance documents in the document database of each word extracted from the seed document as statistical information of this word by starting the statistical information reference program 1700 in step 1800, and referring to the information file 180 for full-text searches.
[0123] In addition, as acquisition of the number of appearance documents of this word was shown from the information file 180 for full-text searches as an information file 803 for full-text searches shown in drawing 8 , it can use that the publication number and appearance location of each word are stored in the information file 180 for full-text searches, and the publication number from which this word differs can be realized by carrying out counting.
[0124] And in step 1801, the significance of each word extracted from the seed document is computed using the statistical information in the count of the appearance in a seed document and document database of this word, and it stores in a work area 170.
[0125] The above is the procedure of word significance calculation program 150a.
[0126] In addition, if it considers as the word significance formula in this example, it is good also as a thing using the TF-IDF (Text Frequency, Inverted Documents Frequency) method, for example.
[0127] The above is the second example of this invention.
[0128] As explained above, the word significance in consideration of the word (it is hereafter called a frequent appearance word) which appears frequently within a document database can be computed now by using the similar document-retrieval system in the second example of this invention. Namely, it is low in the word significance of a frequent appearance word, and by setting up the word significance of a rare word highly, it becomes possible to choose the word showing the description of a seed document preferentially, and a highly precise similar document retrieval can be realized now.
[0129] Next, the third example of this invention is explained using drawing 19 .
[0130] Although the statistical information of the registration document accumulated in the document database is used in case the third example of the similar document-retrieval system which applied this invention computes the significance of the word extracted from the seed document like the second example, it differs in that the statistical information file 1900 is used for acquisition of statistical information.
[0131] According to this approach, the statistical information acquisition referred to in the case of the word significance calculation in the second example can be performed now at a high speed.
[0132] Although this example takes the almost same configuration as the second example (drawing 17), the configurations of the registration control program 111 differ, and as shown in drawing 19 , the statistical information file creation registration program 1900 is added. Moreover, the statistical information file 1910 is stored in a magnetic disk drive 103. At step 1800 of said word significance calculation program 150a, it comes to refer to the statistical information file 1910 shown in drawing 19 instead of referring to the information file 180 for full-text searches, in case the statistical information in the document database of each word extracted from the seed document is acquired.
[0133] Hereafter, the procedure of different registration control program 111a from the second example is explained using drawing 20 .
[0134] In registration control program 111a, in step 1200, the registration document read in program 120 is started first, the document specified as a candidate for registration is read, and it stores in a work area 170.

[0135] Next, in step 1201, the information file creation registration program 121 for full-text searches is started, the information for full-text searches corresponding to the registration document stored in the work area 170 is created, and it stores in the information file 180 for full-text searches.
[0136] Next, in step 2000, the statistical information file creation registration program 1900 is started, the statistical information corresponding to the registration document stored in the work area 170 is created, and it stores in the statistical information file 1910.
[0137] The above is the procedure of the registration control program 111.
[0138] The example of the statistical information file 1910 created by drawing 21 by the statistical information file creation registration program 1900 is shown.
[0139] The management number 2100, a word 2101, and 2102 appearance documents are stored in the statistical information file 1910 shown in this Fig.
[0140] The example shown in this Fig. shows word"LA" being stored in the field of management number"0", and being stored as the number of appearance documents of this word is "1."
[0141] In addition, although the statistical information file 1900 shall be stored by the tabular format in the example shown in drawing 21 , as long as it is the format which can acquire a word and the number of appearance documents, you may be what kind of format. For example, it does not matter as what is stored in a try format, and does not matter as what is stored in the head field of the information file 180 for full-text searches.
[0142] The above is the third example of this invention.
[0143] As explained above, according to the third example of this invention, by referring to the statistical information file beforehand created by acquisition in the statistical information of each word extracted from the seed document at the time of document registration processing, it becomes unnecessary to carry out counting of the number of a different appearance publication number with reference to the information for full-text searches, and statistical information can be acquired now at a high speed. Thereby, a high-speed similar document retrieval can be realized now compared with the second example.
[0144] Next, the fourth example of this invention is explained using drawing 22 .
[0145] The fourth example of the similar document-retrieval system which applied this invention approximates and uses the statistical information of each word extracted from the seed document.
[0146] According to this approach, the capacity of the statistical information stored in the statistical information file 1910 in the third example can be reduced, without reducing the precision of statistical information extremely.
[0147] Although this example takes the almost same configuration as the third example (drawing 19), the configurations of the statistical information reference program 1700 differ, and the approximation statistical information calculation program 2200 is added.
[0148] Hereafter, the procedure of different statistical information reference program 1700b from the third example is explained using drawing 23 .
[0149] Statistical information reference program 1700b repeats and performs steps 2301-2304 about all the words extracted from the seed document (step 2300).
[0150] At step 2301, it checks whether the statistical information corresponding to this word is stored with reference to the statistical information file 1910.
[0151] And when this word is stored in the statistical information file 1910, step 2303 is performed, and step 2304 is performed when not stored (step 2302).
[0152] At step 2303, the statistical information of this word is acquired with reference to the statistical information file 1910.
[0153] Moreover, at step 2304, the approximation statistical information calculation program 2200 is started, and the approximation statistical information of this word is computed.
[0154] The above is the procedure of statistical information reference program 1700b.
[0155] Next, the procedure of the approximation statistical information calculation program 2200 is concretely explained using drawing 24 .
[0156] the word 2400 which serves as an object which acquires statistical information in step 2301 in

the example shown in this Fig. first -- "LAN" -- it receives and the statistical information file 1910 is referred to.
[0157] Here, since "LAN" is not stored in the statistical information file 1910, step 2304 is performed.
[0158] At step 2304, the statistical information of "LA" which is the component of word 2400"LAN", and "AN" is acquired, respectively, and few values are set up as statistical information of "LAN" among these numbers of appearance documents.
[0159] a book -- a Fig. -- having been shown -- an example -- **** -- " -- LA -- " -- statistical information -- 2401 -- storing -- having had -- an appearance -- a document -- a number -- " -- 807 -- " -- "AN" -- statistical information -- 2402 -- storing -- having had -- an appearance -- a document -- a number -- " -- 1512 -- " -- comparing -- this -- a result -- ****** -- " -- LAN -- " -- statistical information -- 2403 -- ****** -- a value -- being small -- " -- LA -- " -- an appearance -- a document -- a number -- " -- 807 -- " -- storing (2410) .
[0160] When these differs in the number of appearance documents of word "component of LAN"" LA", and "AN", the number of appearance documents of "LAN" uses the property in which it cannot increase more than each component. That is, as the number of appearance documents of word"LAN", although the number of appearance documents of the "LAN" itself should be used essentially, it refers to as the number of appearance documents which approximated the value with few appearance documents among "LA" which is the component of word"LAN", or "AN".
[0161] The above is the concrete procedure of the approximation statistical information calculation program 2200.
[0162] The above is the fourth example of this invention.
[0163] Since it becomes unnecessary to store no number of appearance documents of words in a statistical information file by using the similar document-retrieval system in the fourth example of this invention as explained above, the capacity of a statistical information file can be reduced compared with the third example.
[0164] Since the similarity of a seed document is computed in the similar document-retrieval system in the fourth example from the first example of this invention and the number for retrieval of words is adjusted based on this as explained above, a similar document retrieval is realizable for a high speed, securing retrieval precision.
[0165] Next, the fifth example of this invention is explained using drawing 25 .
[0166] The fifth example of the similar document-retrieval system which applied this invention outputs a retrieval result by predetermined retrieval time.
[0167] According to this approach, since a user can acquire a retrieval result by predetermined retrieval time, he can judge without stress whether the seed document specified on retrieval conditions has agreed for the purpose of retrieval.
[0168] Although this example takes the almost same configuration as the first example (drawing 1), the configurations of the similarity calculation program 132 differ and the retrieval processing-time measurement program 2500 is added.
[0169] Hereafter, the procedure of different similarity calculation program 132b from the first example is explained using the PAD diagram of drawing 26 .
[0170] In step 2600, similarity calculation program 132b starts the retrieval processing-time measurement program 2500, and starts measurement of the retrieval processing time.
[0171] Next, if the retrieval processing time becomes below a predetermined value (it is hereafter called the retrieval time limit) to all the words for retrieval stored in the work area 170, steps 1602, 1603, and 2602 will be repeated and performed (step 2601).
[0172] At step 1602, the count acquisition program 161 for retrieval of a word appearance is started, the count of an appearance within each registration document is acquired with reference to the information file 180 for full-text searches corresponding to the word for retrieval, and it stores in a work area 170.
[0173] Next, in step 1603, the similarity calculation program 162 classified by element is started, the similarity according to element of the registration document to a seed document is computed by the predetermined formula using the count of the appearance in a seed document of the word for retrieval

stored in the work area 170, and the count of the appearance in a registration document, and it adds to the similarity of the whole registration document.
[0174] And in step 2602, the retrieval processing-time measurement program 2500 is started, the elapsed time of the retrieval processing time is measured, and the retrieval processing time is computed.
[0175] The above is the procedure of similarity calculation program 132b.
[0176] The above is the fifth operation gestalt of this invention.
[0177] In addition, the retrieval time limit in step 2601 of this example is good also as what is specified as retrieval conditions at the time of retrieval activation, and good also as what is beforehand set up as a system construction value.
[0178] Moreover, since it thinks also when only a small number of word for retrieval is used depending on the set point, you may enable it to set up the minimum number for retrieval of words for maintaining retrieval precision in this example, although the retrieval time limit shall be set up. In this case, even if the retrieval processing time exceeds the retrieval time limit, the specified minimum number for retrieval of words will repeat similar retrieval.
[0179] Furthermore, although the time amount which similarity calculation processing takes using the retrieval processing-time measurement program 2500 shall be measured in this example, it is good also as what measures the retrieval processing itself. In this case, what is necessary is to start the retrieval processing-time measurement program 2500, and just to start measurement of the retrieval processing time, before starting the retrieval condition analyzer 130 by the retrieval control program 112 rather than starting measurement of retrieval time at step 2600 shown in drawing 26 .
[0180] Since the number for retrieval of words is adjusted in the similar document-retrieval system in the fifth example of this invention based on the time amount which retrieval takes as explained above, a retrieval result can be acquired by the predetermined processing time.
[0181] As this result, a user can predict retrieval end time now.
[0182] In addition, it is also possible to use it by the time of retrieval activation or system definition from the first example, changing the similar document-retrieval system which ends retrieval for the retrieval time which explained to the standard the similarity of the seed document explained in the fourth example in the similar document-retrieval system which ends retrieval, and the fifth example to a standard.
[0183] Next, the sixth example of this invention is explained using drawing 27 .
[0184] The sixth example of the similar document-retrieval system which applied this invention asks a user for a check, when presuming retrieval time and requiring huge time amount from the word for retrieval used for retrieval from the word extracted from the seed document.
[0185] According to this approach, by the word extraction condition for retrieval in the similar document-retrieval system explained in the fourth example from the first example, since retrieval can be canceled in advance when retrieval takes huge time amount, a user loses being carelessly kept waiting.
[0186] Although this example takes the almost same configuration as the first example (drawing 1), the configurations of the word extract program 131 for retrieval differ, and as shown in drawing 27 , the retrieval time presumption check program 2700 is added.
[0187] Hereafter, the procedure of different word extract program 131b for retrieval from the first example is explained using the PAD diagram of drawing 28 .
[0188] In the word extract program 131 for retrieval, first, the word significance calculation program 151 is started, the significance of the word stored in the work area 170 based on the predetermined formula is computed in step 1500, and it stores in a work area 170.
[0189] Next, steps 1502-1505 are repeated and performed to all the words stored in the work area 170 at said step 1500 (step 1501).
[0190] First, in step 1502, the word stored in the work area 170 is acquired in descending order of significance.
[0191] Next, in step 1503, the word extract judging program 151 for retrieval is started, and the similarity according to element of a seed document is computed.
[0192] And in step 1504, when the similarity according to element of a seed document judges and is

over whether it is over the predetermined threshold and it is not over step 1505, repeat processing is ended.
[0193] And in step 1505, it stores in a work area 170 by making this word into the word for retrieval.
[0194] Next, in step 2800, when retrieval time is presumed from the word for retrieval stored in the work area 170 and the presumed retrieval time (it is hereafter called presumed retrieval time) exceeds a predetermined value (assignment retrieval time), the message which checks continuation of retrieval is displayed and an user validation is received. As this acknowledgement message, as shown, for example in drawing 6 , the message 2900 which has the continuation carbon button 2901 and Cancel button 2901 may be displayed.
[0195] The above is the procedure of word extract program 131b for retrieval.
[0196] In addition, as assignment retrieval time in the above-mentioned step 2800, it is good also as what is specified as retrieval conditions, good also as what is beforehand specified as system definition, and good also as a certain thing which is, crawls and is automatically set up from the result of the test pattern of shoes.
[0197] Moreover, as the presumed approach of the retrieval time in the above-mentioned step 2800, it is good also as what is presumed from the number of appearance documents of this word for retrieval, and good also as what is presumed from the size of the information file 180 for full-text searches corresponding to this word for retrieval. Or the mean time which one word for retrieval takes using some test patterns is measured, and it is good also as what presumes retrieval time using this mean time.
[0198] Since it becomes possible to adjust the extraction condition of the word for retrieval when retrieval time is presumed from the extracted word for retrieval in the similar document-retrieval system shown in this example and presumed retrieval time exceeds the time amount specified beforehand as explained above, a user loses being carelessly kept waiting.
[0199]
[Effect of the Invention] As explained above, in this invention, the number for retrieval of words which uses the similarity of a seed document for similarity calculation since the number for retrieval of words is set as a standard is reducible. Thereby, the high-speed similar document retrieval which can secure retrieval precision is realizable.

[Translation done.]

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

CLAIMS

[Claim(s)]
[Claim 1] In the similar document-retrieval approach of searching the document with which the contents are similar to the document (it is hereafter called a seed document) specified from the document or text registered into the document database, or the character string (it is hereafter called a document collectively) The index creation step for full-text searches which creates the index for full-text searches of the document made applicable to registration as registration processing of the document to a document database, The seed document feature-vector creation step which creates the vector data (it is hereafter called a seed document feature vector) which used the count of an appearance for every character string contained in the specified seed document as retrieval processing of a similar document as the element, The character string showing the central contents of this seed document to the character string which is the element of said seed document feature vector to that extent The character string extract step for retrieval which follows for (calling it character string significance hereafter), extracts, and extracts the character string (it is hereafter called the character string for retrieval) used for the descending order of this character string significance by predetermined extract criteria at similarity calculation, It is related with the character string for retrieval extracted at said character string extract step for retrieval. The similarity calculation step which computes the similarity of each registration document to a seed document using the appearance information within the seed document of this character string for retrieval, and the appearance information within the document (it is hereafter called a registration document) registered into the document database, The similar document-retrieval approach characterized by having the retrieval result output step which outputs the similarity to the seed document of each registration document computed at said similarity calculation step.
[Claim 2] The similar document-retrieval approach characterized by to have the similarity calculation step which computes the similarity of each registration document to a seed document using the count of an appearance within the seed document of this character string for retrieval, and the count of an appearance within a registration document about the character string for retrieval extracted at said character string extract step for retrieval as said similarity calculation step in the similar document-retrieval approach according to claim 1.
[Claim 3] As said character string extract step for retrieval in the similar document-retrieval approach according to claim 1 The character string significance calculation step which makes the count of an appearance in this seed document the character string significance of this character string about the character string which is the element of the seed document feature vector created at said seed document feature-vector creation step, The similar document-retrieval approach characterized by having the character string judging step for retrieval which extracts the character string for retrieval of the number beforehand specified as the descending order of the character string significance computed at said character string significance calculation step.
[Claim 4] As said character string judging step for retrieval in the similar document-retrieval approach according to claim 3 Instead of extracting the character string for retrieval of the number specified beforehand, the character string used for similarity calculation is extracted in descending order of the

character string significance computed at said character string significance calculation step. The similar document-retrieval approach characterized by using the character string judging step for retrieval which extracts this character string as a character string for retrieval when the similarity to a seed document is computed by this character string and this similarity is over the predetermined value.
[Claim 5] The similar document-retrieval approach carried out [ending similarity calculation processing, when the retrieval processing time measured at the above-mentioned retrieval processing-time measurement step exceeds a predetermined value in said similarity calculation step, while adding the retrieval processing-time measurement step which measures the time amount which retrieval takes as retrieval processing in the similar document-retrieval approach according to claim 1, and] as the description.

[Translation done.]